# Structural characterization of the type II secretion system of *Aeromonas hydrophila*

A Thesis Submitted to the College of

Graduate Studies and Research

In Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy

In the Department of Leisure Activity

University of Saskatchewan

Saskatoon

By

BENJAMIN FLATH

**PERMISSION TO USE**

In presenting this thesis as partial fulfillment of the requirements for a Postgraduate Degree at the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I also agree that permission for copying this thesis in any manner for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Dean of the College in which my thesis work was done. It is understood that any copying, publication, or use of this thesis or parts of this thesis for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis. Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Pharmacy and Nutrition
College of Pharmacy and Nutrition
University of Saskatchewan
110 Science Place
Saskatoon, Saskatchewan
Canada, S7N 5A8

**ABSTRACT**

The *exe*C gene, found in the gram-negative bacteria *Aeromonas hydrophila* codes for a 31 kDa, three domain, bitopic inner membrane protein. The components of the ExeC protein include an amino-terminal cytoplasmic domain, a trans-membrane helix and two periplasmic domains. The two periplasmic domains are involved in recognition and selection of protein substrates which are subsequently transported across the outer membrane and free of the cell. This study focuses exclusively on the two periplasmic domains referred to hereafter as the HR and the PDZ domains. Three constructs were used throughout the course of this study. Two of them were designed, cloned and expressed for this study. The third is a result of previous work. Two constructs contained both the HR and PDZ domains while the other consists of the amino-terminal periplasmic HR domain. Only one construct was used to grow single crystals for analysis by X-ray crystallography. Crystals comprised of the PDZ domain from a degraded construct grew in a hexagonal space group with a hexagonal bi-pyramidal morphology. Crystals diffracted anisotropically to a maximum resolutions of 2 Å along the c axis and 3 Å in the a/b plane. Anisotropy in combination with twinning drastically complicated structure solution. Efforts toward elucidating the crystal structure will be discussed.

**ACKNOWLEDGEMENTS**

There are a number of people I wish to thank. Firstly I would like to thank Dr. Pawel Grochulski who recognised me as a curios and able young scientist eager to get his feet wet in an exciting field of science. Dr. Grochulski's guidance and support was always helpful and sincere. Secondly I would like to thank Dr. Peter Howard for many enthusiastic discussions on laboratory techniques, procedures and general cellular biology. He has been a great tutor.

I would also like to thank the members of my committee, Dr. Jane Alcorn, Dr. Louis Delbaere, Dr. Yu Luo and Dr. Jian Yang. Thanks goes out to Dr. Yang for being very accommodating and generous with regard to the use of his Laboratory. I would like to acknowledge lab mates for their help along the way; Dr. Karin Van Straaten, Wanpeng Sun, Lifeng Chen, Yunyou Su and those in Dr. Howard's lab.

For financial support I would like to thank the Molecular Design Research Group (MDRG), Saskatchewan Health Research Foundation (SHRF), the College of Pharmacy and Nutrition at the University of Saskatchewan and the John Delfari Research Trust.

Lastly but certainly not least I would like to thank my wife Carli for her support and good spirited involuntary learning of crystallography. My apologies go out to my beautiful daughter Violet for she has no defense against crystallographic infusion. I would also like to mention the rest of my family; Mom, Dad, Brook, JoAnne, Jessica, Jordon, Katelyn, Riley, Brett, Darcy, Cheryl and Elgin.

**TABLE OF CONTENTS**_Toc346291069

**LIST OF TABLES**

LIST OF FIGURES

## LIST OF ABBREVIATIONS

**Table 1.0**

| Abbreviation | Description |
|---|---|
| AMP | Adenosine monophosphate |
| CMCF | Canadian Macromolecular Crystallography Facility |
| DLS | Dynamic Light Scattering |
| DNA | Deoxyribonucleic |
| EDTA | Ethylenediaminetetraacetic Acid |
| GSP | General Secretion Pathway |
| IM | Inner Membrane |
| IMAC | Immobilized Metal Affinity Chromatography |
| IPTG | Isopropyl-Beta-D-Thiogalactopyranoside |
| JCSG | Joint Center for Structural Genomics |
| MAD | Multi-Wavelength Anomalous Dispersion |
| MR | Molecular Replacement |
| MS | Mass Spectrometry |
| MTB | Main Terminal Branch |
| MX | Macromolecular Crystallography |
| NCS | Non-Crystallographic Symmetry |
| NMR | Nuclear Magnetic Resonance |
| NTA | Nitrilotriacetic Acid |
| NTP | Nucleotide Tri-Phosphate |
| OM | Outer Membrane |
| PCR | Polymerase Chain Reaction |
| PDB | Protein Data Bank |
| PEG | Poly Ethylene Glycol |
| PG | Peptidoglycan |
| PMP | p-methoxyphenyl |
| PMSF | phenylmethylsulfonyl fluoride |

| | |
|---|---|
| SAD | Single-Wavelength Anomalous Dispersion |
| SAXS | Small Angle X-ray Scattering |
| SDS-PAGE | Sodium-Dodecyl-Sulphate Poly-Acrylamide Gel Electrophoresis |
| SeMet | Seleno-Methionine |
| SSSC | Saskatchewan Structural Science Centre |
| SSSC | Saskatchewan Structural Science Centre |
| T2SS | Type II Secretion System |
| TCEP | Tris(2-carboxyethyl)phosphine |
| TM | Trans-Membrane |
| XAS | X-ray Absorption Spectroscopy |

# 1.0 HYPOTHESIS

All proteins within a living organism have specific functions which are directly related to their structure. Therefore in order to better understand macromolecular mechanisms like the Type II secretion system (T2SS) one must understand the underlying structure. It is hypothesized that structural characterization of the T2SS using X-ray crystallography and associated techniques will help understand their function, specifically that of the ExeC protein belonging to the T2SS of *A. hydrophila*.

# 2.0 OBJECTIVES

1) Design an ExeC construct that can be produced in large amounts and readily purified in order to facilitate further analysis.
2) Crystallization of the purified ExeC protein, which is a prerequisite to X-ray crystallography.
3) Solve the three dimensional atomic structure of the ExeC construct to further understand its role in the T2SS.

# 3.0 INTRODUCTION

## 3.1 The Type II Secretion System

Gram negative bacteria are unique to other cell types in that they have two phospholipid bi-layers surrounding them. They have an inner membrane (IM) also known as the cytoplasmic membrane and an outer membrane (OM). The space between these two layers is referred to as the periplasm which contains peptidoglycan (PG). The periplasmic space is unique to gram negative bacteria and therefore these bacterial cells have evolved unique mechanisms of transporting material across this double layer. Sophisticated protein assemblies have evolved to selectively export proteins across both the inner and outer membranes and into the extracellular milieu. One such mechanism is the type II secretion system (T2SS), alternatively referred to as the main terminal branch (MTB) of the general secretion pathway (GSP). The T2SS is also an integral part of type IV pili formation which is responsible for twitch motility (Hazes and Frost 2008).

The T2SS was first found in *A. hydrophila* by Howard *et al.* (Jiang and Howard 1992; Howard *et al.* 1993) First, *exe*E and *exe*F were found through mutational studies where secretion was hampered and later reconstituted by complementation. This led to further analyses where the T2SS operon was further characterized.

Secretion of exoproteins happens in two stages. Newly synthesized proteins destined for translocation across the IM are produced with an amino-terminal signal sequence. This signal sequence is recognized by the *sec* translocation machinery which inserts the chaperoned protein into the IM and cleaves off the signal peptide. Once inserted in the IM the protein is effectively folded and some are post-translationally modified and/or oligomerized (Pugsley 1993). Secondly, depending on species, 12-15 unique gene products assemble to form the mega-Dalton secretion machinery. The configuration of these proteins within the T2SS is shown below (Figure 3.1).

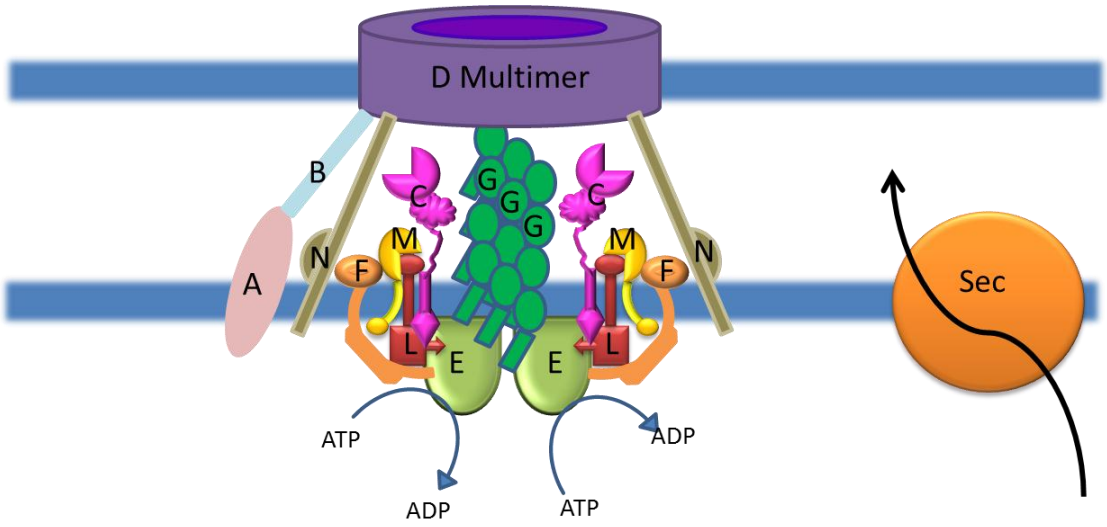



**Figure 3.1 Type II Secretion System Diagram**

The T2SS is shown here as a schematic diagram, illustrating the relationship between the components of the T2SS discussed above. GspE (light green) is an ATPase that has a hexameric ring and is held in association with the IM via its interaction with GspL (red). GspL interact with GspM (yellow) and is reported to be important for cell localization of the secretion system. GspF (light orange) is also a mostly helical bitopic protein necessary for functional secretion. Another bitopic inner membrane protein is GspC (magenta) which is believed to act as a gatekeeper of the secretion system. GspN (olive) is not found in all T2SS and is poorly characterized. It is known to also be a bitopic IM protein which interacts with GspD (purple). GspD is imbedded in the outer membrane and forms the opening through the OM which allows the fully folded protein to be secreted out of the cell. GspA (pink) has a nucleotide binding site as well as interact with peptidoglycan. GspA is always found with GspB (light blue) and in some organisms GspA-B are the same gene products. The pseudopilins denoted above as ‘G’ (the major pseudopilin) have several genes coding for additional minor pseudopilins H, I, J and K. GspG-K form a pilin like structure that is capable of assembly and disassembly which is believed to fundamental to secretion regulation. The large orang circle denotes an entire system called the Sec system which uses amino terminal signal sequence on proteins that are destined for the periplasmic space.

**General Secretion Pathway protein E (GspE)**

The inner membrane complex is where the vast majority of the secretion machinery resides and at the core of the assembly is a nucleotide triphosphate. A wealth of literature is accumulating on GspE and its family of 'type II /IV secretion NTPases'. GspE belongs to a large superfamily of NTPases. Various portions of the homologue from *Vibrio cholerae* have been elucidated by X-ray crystallography (Robien *et al*. 2003; Abendroth *et al*. 2005) (Figure 3.2). The GspE homologue from *Vibrio cholerae,* denoted EpsE, was solved by Robien *et al.* The GspE homolouge from *Archaeoglobus fulgidus* was solved at just better than 3 Å resolution in space group R32 with two monomers per asymmetric unit. AMP-PNP bound and unbound states were observed. Interstingly, the observation of molecules related by a 3-fold symmetry axis recurrs in the the hexameric structure of GspE, wich was observed in small angle X-ray scattering (SAXS) solution studies (Yamagata and Tainer 2007). The same authors also elucidated the most complete structure of GspE and it is only missing 90 residues from the amino-terminal end (Figure 3.3). GspE is a multi-domain protein. There are two distinct domains; the C-terminal and N-terminal domains which are further subcategorized. The carboxyl-terminal domain consists of three subdomains denoted as C1, C2 and Cm. The C1 subdomain is the largest of the three subdomains and contains the NTP binding site with identifiable Walker A motif or phosphate-binding loop (P-loop) and Walker B motif, which coordinates the $Mg^{2+}$ essential to ATP hydrolysis. The C1 subdomain has a central β-sheet that consists of six parallel strands and a seventh antiparallel strand with five helices surrounding the β-sheet.

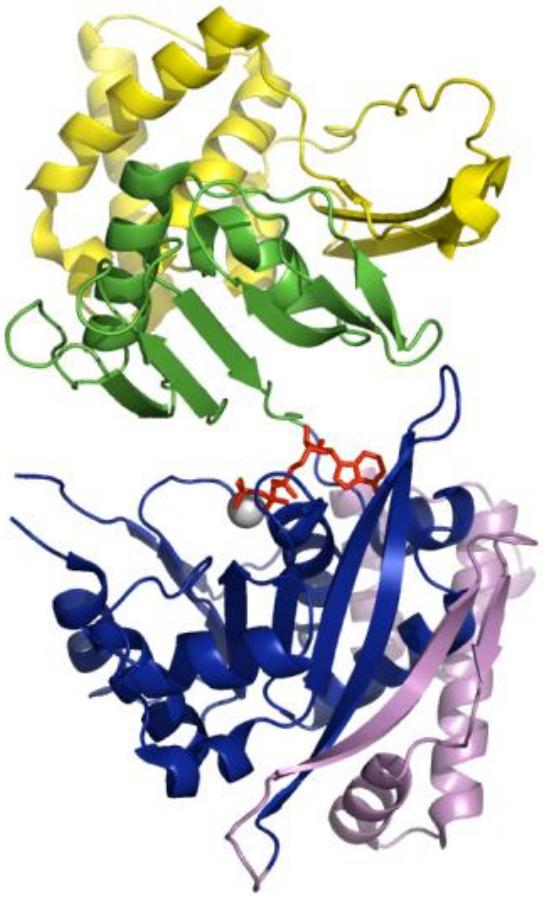


**Figure 3.2 GspE Structure**

The ribbon diagram of a monomer of GspE viewed from front and back. The yellow N1 and green N2 domains have hydrophobic patches that associate with the inner part of the cytoplasmic membrane. The blue C1 domain and the magenta C2 domain is where the nucleotide binding site and $Mg^{2+}$ (grey sphere) binding site reside in the active site PDB code: 2OAQ (Yamagata and Tainer 2007) (*PYMOL*).

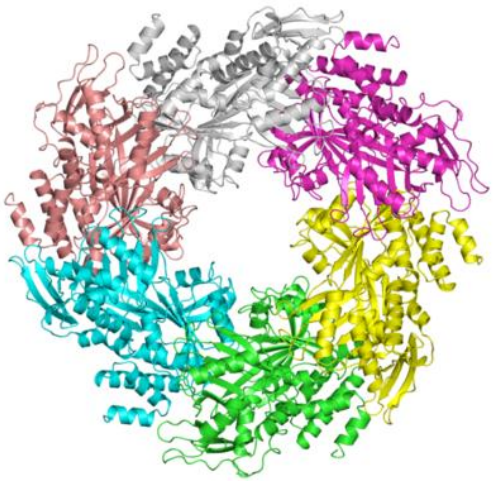


**Figure 3.3 Hexameric GspE Structure**

The hexameric structure of GspE forming a ring structure with a 25 Å pore. Each chain is shown in a different colour. Notice each chain does not have an identical conformation in fact the hexamer is more accurately described as a trimer of dimers. PDB code: 2OAP (Yamagata and Tainer 2007) (*PYMOL*).

---

The C2 subdomain is made up of four small helical pieces that make up the terminal 50 residues of the protein. C2 lies between the interface of the C1 and Cm subdomains and has a number of highly conserved contacts with the Cm domain. The Cm subdomain's primary feature is a conserved tetra-cysteine motif that coordinates an unknown divalent metal ion believed to be a Zn ion (Robien *et al.* 2003). This however remains to be distinguished, possibly using X-ray absorption spectroscopy (XAS).

The amino terminal domain is also subdivided. The N-terminal domain has been solved, both as part of the whole GspE protein but as well it has been solved as a fragment that is complexed with GspL. Interactions with GspL have been found to be necessary for adherence to the inner side of the cytoplasmic membrane. It is interesting to note that GspE has been shown to exist as a functional hexamer. Also the sequence conservation between the closest homologues shows that in its folded form the residues nearest to the IM are most highly conserved while those residues further away from the membrane and in the cytosol are more variable (Robien *et al.* 2003).

Supporting the hexameric assembly of GspE is the fact that several ATPases from the type IV secretion family have also been characterized as hexamers (Johnson *et al*. 2006).

**General Secretion Pathway protein L (GspL)**

Like many members of the inner membrane complex, GspL is a bitopic IM protein consisting of ~400 residues. The largest domain of GspL is the cytoplasmic domain consisting of 250 residues which is recruited and held in place by GspE against the inner membrane. A single trans-membrane helix links the cytoplasmic domain to the 130 amino acid periplasmic domain. GspL from *Aeromonas hydrophila* has 36% sequence identity to that of its *Vibrio cholera* counterpart. Cyto-GspL from *Vibrio cholera* was found to complex with the N-terminal domain of GspE. The crystal structure of this complex forms a hetero-tetramer with a surface area of 1700 $Å^2$ buried from the solvent as a result of the interaction between GspE and GspL monomers. The central dimer of the complex consist of two GspL cytoplasmic domains and has been found in two unrelated crystal structures implying a physiological relevance (Figure 3.4) (Abendroth et al. 2005), (Abendroth *et al.* 2009). The structure of the periplasmic domain of GspL has also been solved (Abendroth *et al.* 2004).

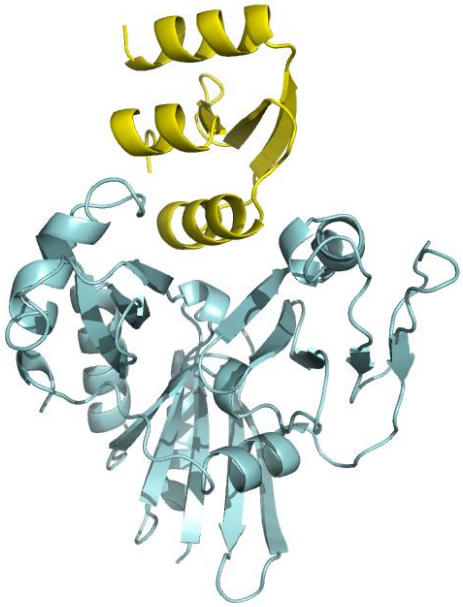


**Figure 3.4 GspE-L Complex**

The GspE-L complex was solved to 2.4 Å and consists of cyto EpsL domains I, II, and III bound to N1 (yellow) of EpsE (GspE). EpsE 1-96 and EpsL 1-241 are the first structural representation of any two proteins from the T2SS bound to one another (Abendroth *et al.* 2005). The Eps notation is specific to *Vibrio cholera* and can be considered equivalent to Gsp PDB code: 1YF5 (P*YMOL*).

---

The structure of the periplasmic GspL dimer displays close structural similarity to GspM (Figure 3.5). Two regions of EpsL are involved in species specific protein-protein interactions with GspE while the TM helix and/or the periplasmic domain of GspL interacts with EpsM in the GSP of *Vibrio cholera* (Sandkvist *et al.* 2000). Additional interactions have been observed between GspL and the pseudopilin GspJ (Douet *et al.* 2004).

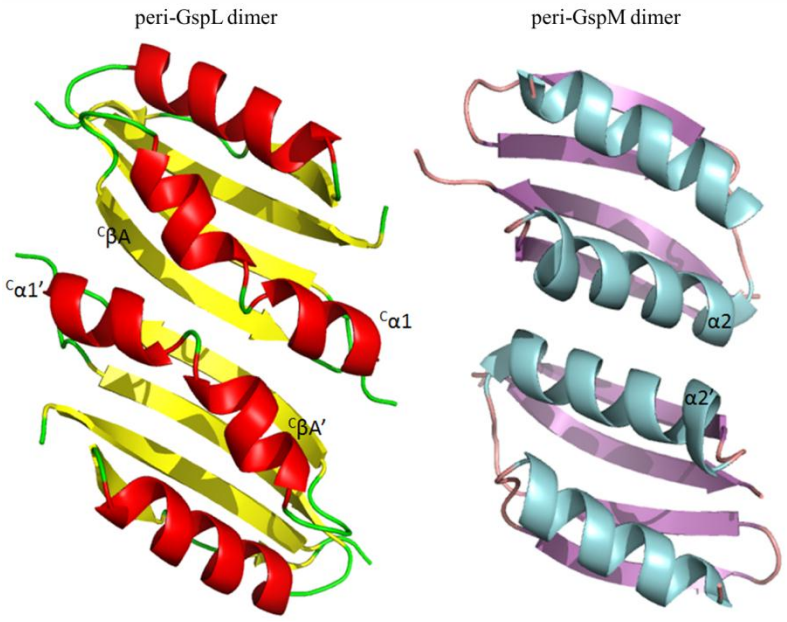



**Figure 3.5 GspL and GspM Dimer**

The similarity in architecture is clearly observed between EpsL and EpsM; both dimers are related by a 2-fold axis and have the same fold. The interaction between dimers is different. The EpsL dimer is mediated by α1- α1' and also βA-βA' (PDB code: 2W7V). EpsM relies on α2-α2' interactions to mediated dimerization (Abendroth *et al.* 2009) (PDB code: 1UV7). (*PYMOL*)

---

## General Secretion Pathway protein M (GspM)

GspM is another bitopic inner membrane protein that consists of ~165 residues. The first 23 residues of the protein are predicted to be in the cytoplasm followed by a predicted TM helix. The majority of the protein is found in the periplasm.

The structure of peri-GspM has been solved and appears to have a ferridoxin type fold (Abendroth *et al.* 2004) which has significant similarities to GspL (Figure 3.5). The fold consists of four beta sheets and two alpha helices in an αββ - αββ conformation. It was shown that fluorescently tagged EpsC and EpsM are localized along the length of the cell (Abendroth,

Kreger *et al.* 2009). This is somewhat contradictory to the reports of others who found that GspM is important for localization of the T2SS to the poles of the cell (Scott *et al.* 2001).

**General Secretion Pathway protein F (GspF)**

The structure of the N-terminal cytoplasmic domain of EpsF was solved by Se-SAD phasing to 1.7 Å (Abendroth *et al.* 2009). This domain was found to form a tight dimer (Figure 3.6). The interacting residues are highly conserved among close homologues indicating physiological relevance. Each domain is composed of a six antiparallel helix bundle. This protein is found to be absolutely indispensable to protein secretion (Sandkvist *et al.* 1997).

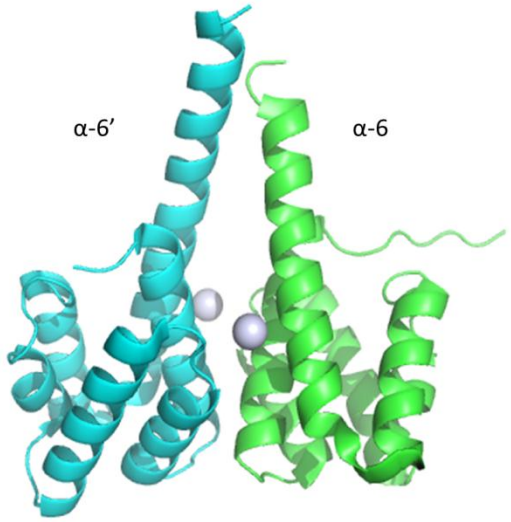



**Figure 3.6 GspF**

The N-terminal cytoplasmic domain of EpsF from residues 56-171. Ligands form both monomers of the EpsF dimer contribute to the coordination of the Ca ion (grey sphere). α-6 of EpsF protrudes significantly from the rest of the assembly. Secondary structure prediction indicates that a trans-membrane helix may extend through to the membrane (Abendroth *et al.* 2009) PDB code: 3C1Q (*PYMOL*).

**General Secretion Pathway protein N (GspN)**

Comparatively little is known about GspN, as it is not found among many type II secretion species. It has been identified in *Klebsiella Oxytoca, Erwinia carotovora, Xanthomonas Campestris, Aeromonas Hydrophila* and *Pseudomonas Putida*. In the case of *K. Oxytoca,* GspN was not required for functional secretion (Possot *et al.* 2000). It has been found through co-precipitation experiments to associate with the porin GspD (Lee *et al.* 2000). GspN is also a bitopic inner membrane protein like GspC.

**General Secretion Pathway proteins A & B (GspAB)**

ExeA and ExeB are somewhat less ubiquitous among T2SS containing species. Through sequence analysis, GspA has been found to have a nucleotide binding site. GspA also interacts with peptidoglycan. These two proteins are essential to the secretion pathway of *A. hydrophila* (Schoenhofen *et al.* 1998). GspA and GspB form a complex with the IM which is proposed to provide energy to the secretion system (Filloux 2004). The ExeAB complex plays a critical role for the localization and assembly of the porin (Ast *et al.* 2002).

**Pseudopilins**

Three dimensional structures for pseudopilins G, H, I, J, K have been solved in recent years (Forest 2008; Korotkov and Hol 2008; Yanez *et al.* 2008; Douzi *et al.* 2009; Korotkov *et al.* 2009; Lam *et al.* 2009). GspG is considered the major pseudopilin protein. It is found in the highest abundance compared to the other pseudopilins (Korotkov and Hol 2008) (Figure 3.7).

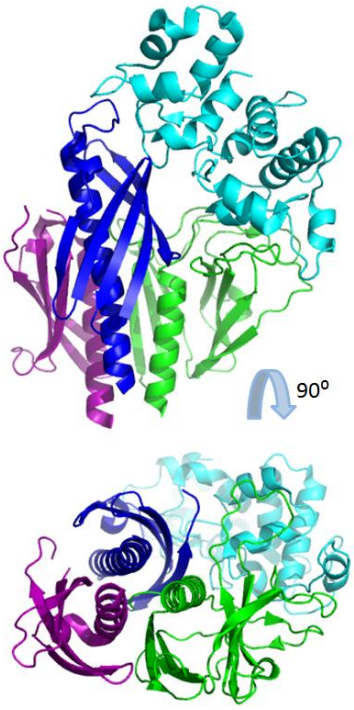



**Figure 3.7 Gsp I, J, K**

The crystal structure of complexed pseudopilins Gsp I,Jand K in complex. Gsp I,J and K are three of the minor pseudopilins. Purple is GspI green is GspJ. GspK has two domains the α-domain cyan and the pilin domain is blue (Korotkov and Hol 2008) PDB code: 3CI0 (*PYMOL*).

---

**General Secretion Pathway protein D (GspD)**

GspD forms the outer membrane pore known as the secretin. This complex consists of 12-14 monomers which are 50-70 kDa in size. They assemble to form an opening in the OM. The C-terminal 300-400 residues consist of the most highly conserved part of the protein whereas the N-terminal domain is more variable. It is through this portion of the secretin that interactions occur between other components of the T2SS such as GspC and, tentatively, GspN. It is through this pore that fully folded proteins move from the periplasm to the extracellular milieu. One example is aerolysin in the case of *A. hydrophila*. There is a necessity for this pore whose opening is somewhere between 50 and 90 Å in diameter to be regulated.

The fully assembled porin appears in most cases to have 12 fold symmetry (Brok *et al.* 1999) sometimes described as a tetramer of trimers (Collins *et al.* 2004). The structure of the N-

terminal domain of GspD from Enterotoxigenic *Escherichia coli* has been solved at 2.8 Å and reveals the structure of the first 237 residues corresponding to the periplasmic portion of the protein (Figure 3.8) (Korotkov *et al*. 2009).

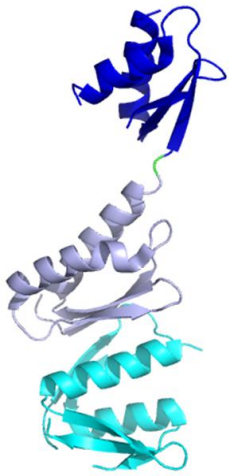


**Figure 3.8 GspD**

The structure of the periplasmic domain of GspD from Enterotoxigenic *E. coli*. This structure consists of the N0, N1 and N2 domains, colored in cyan, light blue and dark blue respectively. The authors state that the β2 strand at the bottom of the figure may play an important role in interacting with other T2SS components. The C-terminus is believed to reside closest to the OM (Korotkov *et al.* 2009) PDB code: 3EZJ (*PYMOL*).

---

**ExeC**

ExeC (the GspC homologue in *A. hydrophila*) is arguably one of the most critical proteins in the T2SS. It is believed to have a regulatory role within the T2SS. It is for this reason that it has been the focus of intensive study including the one presented here. GspC is a bitopic inner membrane protein consisting of a small cytoplasmic domain followed by a single TM helix. Immediately following the TM helix is a 50 residue linker peptide. Two periplasmic domains follow, the first is the HR domain and the second is the PDZ domain which in some species is replaced by a coiled coil domain. The cytoplasmic domain has been shown to interact with the GspL-GspM hetero-multimer, forming a stable complex consisting of GspF-GspL-GspM-GspC (Robert *et al.*

2002; Tsai *et al.* 2002). The sequence alignment for ExeC and its closest homologues are shown below (Figure 3.9).

Interactions between GspC and GspD have been found to be very specific, as complementation experiments require that these two proteins come from the same organism, whereas cross species complementation can be successful using other T2SS components (Bleves *et al.* 1999). GspC and GspD are most often found adjacent to each other in T2SS operons. In the case of *P. Aeruginosa* they are localized in a single operon (Akrim *et al.* 1993). Akrim *et al*. also demonstrated that the trans- membrane domain of GspC is important because mutational studies that replace the domain with that of TetA T2SS results in loss of functionality.

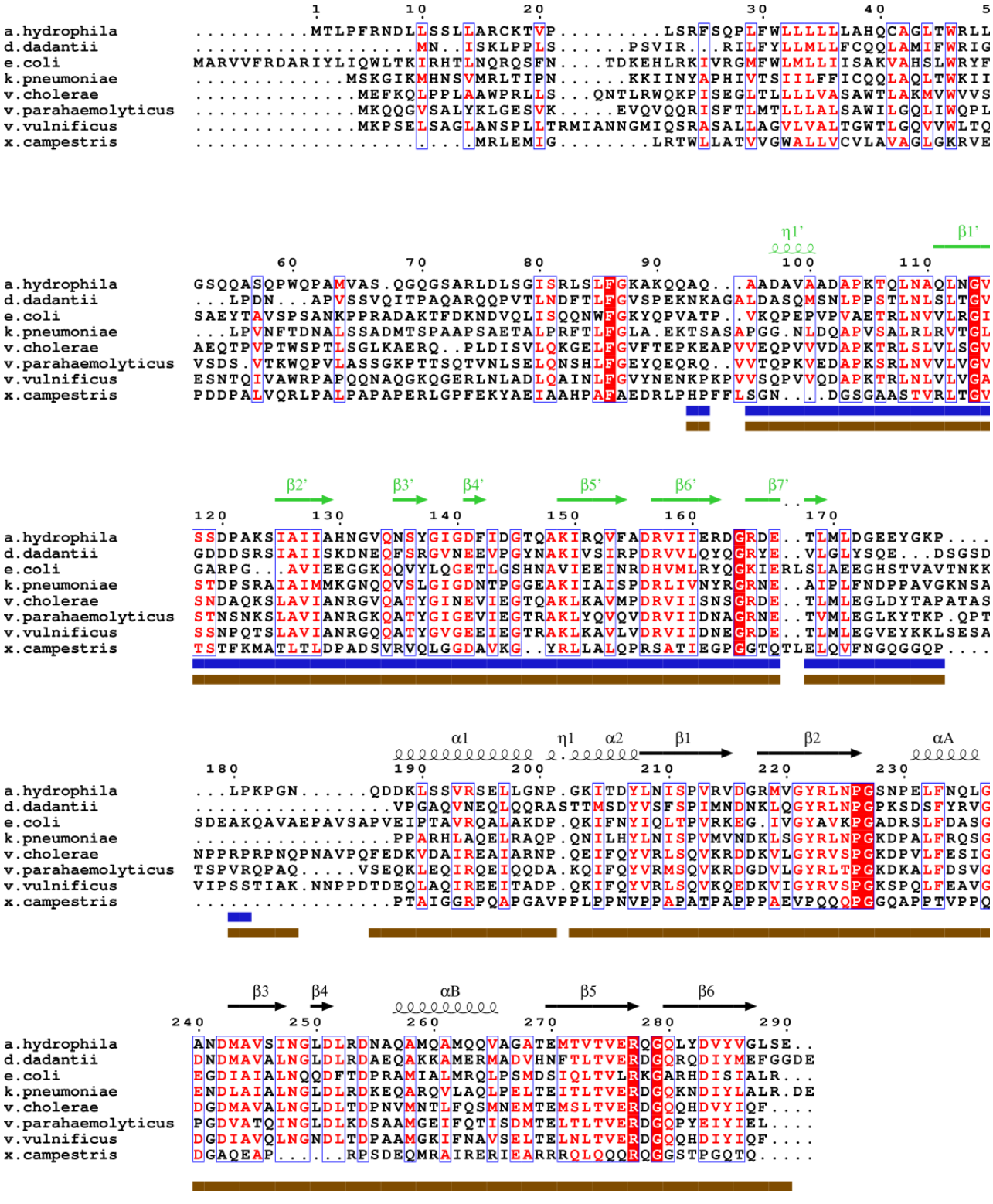



**Figure 3.9** (description on next page)

**Figure 3.9 GspC Multiple Sequence Alignment**

Sequence alignment of ExeC with GspC proteins from other closely related bacterial species. The secondary structure elements for the HR domain (green) are from PDB 2LNV (*Dickeya dadantii*). Secondary structure elements for the PDZ domain (black) are from PDB 2I4S (*Vibrio cholerae*). Accession codes are as follows: *Aeromonas hydrophila* CAA47125, *Dickeya dadantii* Q01564, *E. coli* YP_001464425, *Klebsiella pneumoniae* AAA25125, *Vibrio cholerae* AAA58784, *Vibrio parahaemolyticus* ZP_05119400, *Vibrio vulnificus* NP_933006, *Xanthomonas campestris* AAA27614. The fragments made for this study are shown below the sequence alignment in blue (ExeC-HR) and brown (ExeC-HR-PDZ). The alignment was made with clustalW2 (Chenna *et al.* 2003) using the Gonnet scoring matrix (gap penalty 10, gap extension penalty 0.2) visualized with ESPript 2.2 ESPript (Gouet, P. *et al.* 1999). ESPript: multiple sequence alignments in PostScript (Waterhouse *et al.* 2009).

### 3.1.1 HR Domain

The HR domain was one of the last structurally characterized components of the T2SS. This domain gets its name because of the high homology among varying species, hence, Homologous Region (HR). The GspC homologue of *P. aeruginosa* is an interesting example, where the C-terminal coiled-coil domain can be deleted without compromising the functional secretion system (Bleves *et al.* 1999). In a study conducted in 2006, the researchers also conclude that the HR domain of GspC interacts with GspD (Korotkov *et al.* 2006). This finding was later supported when the crystal structure of the HR domain in complex with N0 and N1 domains Gsp D (Korotkov *et al*. 2011). The β1 strand from HR-GspD and the β1 strand from GspD-N0 domain were involved in the interaction. A structural alignment with the NMR solution structure of Gsp-HR (Gu *et al*. 2012) agrees well with the crystal structure of GspC-HR.

### 3.1.2 PDZ Domain

The PDZ class of proteins are widespread and found in all types of living organisms. The PDZ domain was first characterized as an approximately 90 residue repeat sequence in PSD-95/SAP90, later also found in the Discs-Large Drosophila septate junction protein and the ZO-1

epithelial tight junction protein. This is where the name PDZ originates. The domain is usually about 90 residues long consisting of six β-sheets and two α-helices. The domains are well-known to be a protein interacting domain with a common binding groove between βA and αB of the domain. PDZ domains most often augment the exposed βA strand so that the ligand join the β-sheet by adding another beta strand alongside βA and is held in the binding pocket created between αB and βA. When comparing PDZ domains from different organisms the relative position and length of the αB is variable. The PDZ domain of the T2SS is believed to bind the secreted protein in some fashion.

β1 and β2 form a loop where a highly conserved arginine or lysine preferentially bind to the free carboxylate group at the end of the peptide ligand with a GLGF motif. The family of GspC PDZ domains do not have an identifiable GLGF motif. In addition one of the closest homologues, *V. cholerae,* has been observed to crystallize with α-helix occupying the binding pocket between αB and βA. The evidence suggests that the probability of a binding mechanism alternative to the carboxylate GLGF model is high (Figure 3.10).

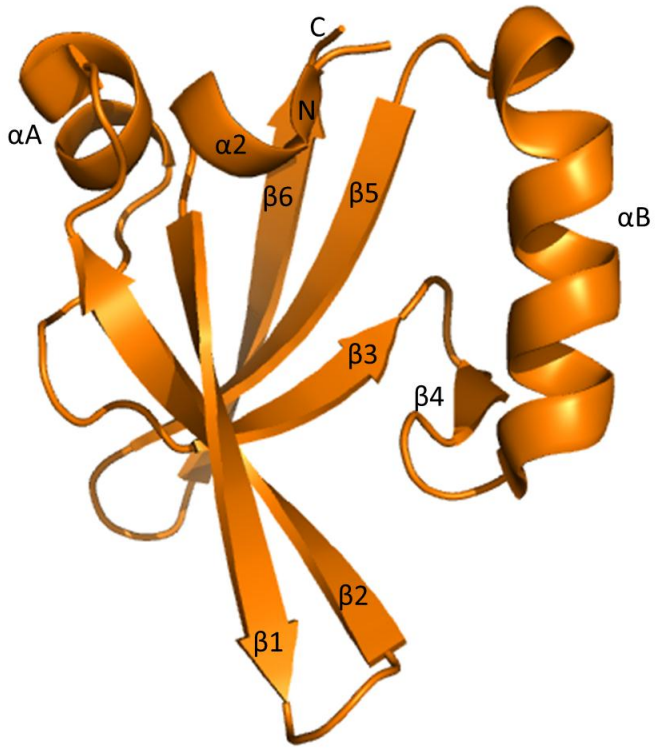



**Figure 3.10 GspC**

The PDZ domain of GspC from *V. cholera* has been solved to 1.63 Å. The domain N-terminal starts with the short α2 helix the next 20 amino acids form β1 and β2 which form a two stranded twisted anti-parallel β-sheet. β1 is surface exposed while β2 is on the protein interior. A small helix αA is part of the linking peptide between β-sheet β1-β2 and β-sheet β3-β4-β5-β6. After αA strand β3 can be found in the centre of the protein. B4 forms a one residue beta sheet before 8 residue αB extends away from the core of the protein to form the peptide binding pocket. The C-terminus of αB ends in a two anti-parallel β-sheet structure composed of strands β5-β6 which joins back with β3 in an anti-parallel manner. The structure shown here has a large binding pocket compared to other PDZ domains. This protein was found to crystalize with αA from a symmetry related molecule embedded in the binding pocket. An α-helix is much bulkier than the β-sheet style interaction with the ligand and β1 found with most other PDZ domains (Korotkov *et al.* 2006). PDB code: 2I6V (*PYMOL*).

## 3.2 Macromolecular X-ray Crystallography

The first ever recorded X-ray diffraction experiment was conducted by Max von Laue in 1912 for which he won the Nobel Prize two years later. These experiments were performed on simple crystals such as diamond, rock salt and zinc sulfide. Although it was known at that time that proteins could be crystallized, the first protein crystal diffraction pattern would not be recorded until three decades later. In 1957 Perutz and Kendrew solved the first protein crystal structure. There was no protein data bank at the time and therefore no protein data bank (PDB) entry represents the initial model. However, in 1971 the PDB was established and in 1973 Watson and Kendrew deposited 1MBN, a 2.0 Å structure of sperm whale myoglobin refined with then-current procedures.

### 3.2.1 Protein Purification and Crystallization

Protein purity is of paramount importance with respect protein crystallization. Contaminants such as dust, peptides, alternate confirmations of the same protein or trace amounts of contaminating proteins will likely start to cause the crystal lattice planes to misalign resulting in increased mosaicity and even cessation of growth (Figure 3.11).

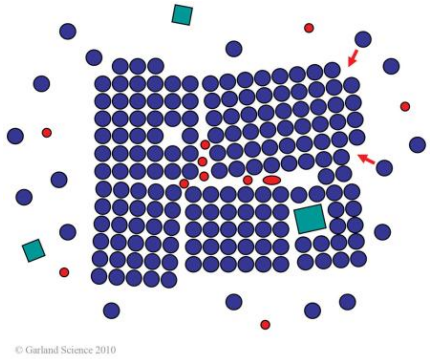



**Figure 3.11 Crystal Mosaicity**

This is an example of a highly mosaic crystal with a number of contaminants present causing lattice dislocations and misalignments (From Biomolecular Crystallography by Bernhard Rupp, © 2012 Garland Science/Taylor & Francis LLC).

---

Crystallization of macromolecules is often the limiting step in crystallography. Crystals can be defined as the periodic assembly of identical building blocks that form a unit cell that can be stacked upon each other in three dimensions forming defined lattice geometry. Because proteins and DNA are very large and irregularly shaped, it is a wonder they can be crystallized at all. It should be noted that for protein crystallography high quality single crystals must be generated if they are to be used for structure solution. Protein crystals are delicate owing to their high solvent content and relatively weak intermolecular interactions. Crystal contacts are those intermolecular contacts that occur between symmetry related molecules from neighboring unit cells within the crystal. These contacts are usually relatively weak interactions. The most common crystal contacts are maintained by; dipole-dipole interactions, hydrogen bonds, salt bridges, van der Walls contacts or hydrophobic interactions. Not only are these weak interactions but there are often very few of them. The key to growing any crystals is to achieve supersaturation. This can be done by adding precipitant, removing water, changing the solvent, temperature or the pH. As these variables change to a condition where the surface of the protein is less compatible with the solution, it may begin to aggregate. A hypothetical solubility diagram and describes what happens as the protein to precipitant ratios change (Figure 3.12).

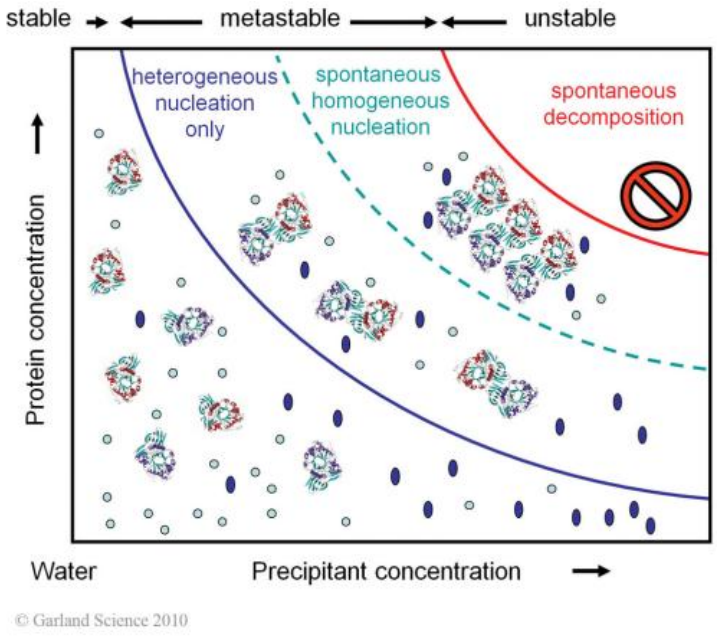



**Figure 3.12 Solubility Diagram**

Dark blue ovals are precipitant. Light blue circle represent water. Three distinct zones are illustrated. At high protein and precipitant concentration is the zone where spontaneous decomposition occurs. The protein readily aggregates and often becomes denatured. The metastable zone is where the protein remains folded but is also supersaturated. Crystal nucleation occurs in the upper portion of the metastable zone while in the lower part of the meta stable zone little or no nuclei are formed however a crystal can continue to grow. The stable zone is where the protein in not supersaturated, no nucleation or crystal growth will occur (From Biomolecular Crystallography by Bernhard Rupp, © 2012 Garland Science/Taylor & Francis LLC).

### 3.2.2 Cryoprotection

Cryoprotection is necessary when data is to be collected at cryogenic temperatures in order to reduce radiation damage. Because about half of a protein's crystal volume is solvent, when it is subject to changing temperature such as cooling, the thermal expansion of the solvent can adversely affect the crystal packing. The solvent surrounding the crystal in the loop can also cause a crystal to bend and deform. When water freezes it forms hexagonal ice which is less dense than liquid water. To avoid this one must add chemical substances to the solvent so that

when frozen the water does not form structured ice but becomes vitrified. The faster one can cool the crystal the less time the solvent molecules will have to order themselves. Very small volumes of water can be flash cooled and converted to a vitrified state, but most often for a crystal in a loop a cryoprotectant is required.

### 3.2.3 X-ray Diffraction from Crystals

Crystals consist of a collection of geometrically regular, periodically repeating units of identical size and shape. The unit cell is the primary voxel that is representative of the entire crystal by way of periodic translations (Figure 3.11). Crystals can form in one of the seven crystal systems; triclinic, monoclinic, orthorhombic, tetragonal, rhombohedral, hexagonal or cubic. The crystal system defines the geometry of the unit cell. The repeating arrangement of points that results from an array of ordered unit cells generates the crystal lattice. Additionally, lattice centering can occur in monoclinic, orthorhombic, tetragonal and cubic lattices. Together the seven crystal systems and the lattice centered counterparts generate the fourteen Bravais lattices.

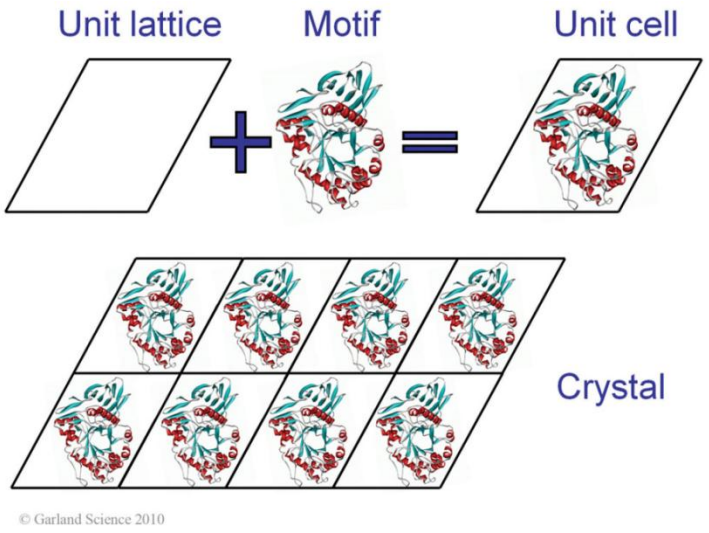



**Figure 3.13 Schematic of a Crystal**

A typical crystal can be thought of as a three-dimensional lattice defined by the regular packing of a particular motif, in this case proteins. It is this quasi-infinite array of translated unit cells that give rise to the diffractive properties of crystals (From Biomolecular Crystallography by Bernhard Rupp, © 2012 Garland Science/Taylor & Francis LLC).

Point group symmetry describes how a single motif rotated around one or more principal axes can generate the 32 crystallographic point groups. A basic example is the radial arrangement of petals on a flower. Point group operators are of the form 360°/*n*. In the three dimensional crystallographic environment only 1, 2, 3, 4 and 6 fold point group rotations can occur while still satisfying the Bravais lattice requirements. Combining the thirty-two point groups with the fourteen Bravais lattices and allowing for symmetry elements such as translations, screw axes, and glide planes result in 230 possible space groups. Protein molecules are chiral and if inversion or mirror symmetry operators were applied the symmetry related copy would have the opposite handedness, which is impossible, as only L-amino acids are normally found in nature. Taking this into account, proteins are only found to crystallize in 65 of the 230 space groups.

The regular packing of a crystal is the reason why they can be used to gain structural information on miniscule samples such as proteins. Because millions of protein molecules are packed together the interaction between the photons from the X-ray beam and the electrons belonging to the protein molecules is amplified.

X-ray scattering occurs when the electrons associated with a molecule are forced to oscillate in response to the electric field of the incident wave. This oscillation produces waves that interfere with each other giving rise to the observed X-ray diffraction pattern. The diffraction of X-rays was first described mathematically by W.L. Bragg, Resulting in Bragg's Law which illustrates that reflections are only observed when scattering of X-rays from electrons happens in phase as shown by the following equation.

$$2d_{\mathrm{hkl}} \sin\theta = \mathrm{n}\,\lambda$$

What Bragg's Law describes is the constructive interference of scattered X-rays which occurs when an equidistant lattice spacing *d* is an integer of the source X-ray wavelength λ. The angle θ of the scattered wave λ is related to the interplanar distance $d_{hkl}$ for the set of reflected planes *hkl*. The scattering observed is dependent on the location and the number of electrons (and therefore the types of atoms) within the reflecting plane. In the context of crystallography, scattering of X-rays by electrons is described as elastic scattering. Elastic scattering, also known as Thomson

scattering, occurs without energy loss and therefore the incoming photon and emitted photon have the same energy. It is the electric field vector of the photon that induces the oscillation of the electron. The oscillating electron will then produce radiation of the same frequency in a mostly forward direction. Thomson's formula describes the intensity of coherently scattered X-rays from a charged particle.

$$I_c = I_0 \frac{e^4}{m^2 r^2 c^4} \left( \frac{1 + cos^2 2\theta}{2} \right)$$

The equation gives the intensity of the coherently scattered X-ray $I_c$, $I_0$ is the incoming intensity of the photon beam and $e$ is the charge, $m$ is the mass, $r$ is the diameter of the charged particle and $c$ the speed of light. The second part of the equation is the polarization factor. Scattering is weakest perpendicular to the incoming beam and strongest in the forward and backward directions. In general X-ray scattering from a crystal weakens with increased scattering angle and also with increasing disorder.

The scattering of X-rays from a quasi-infinite array of unit cells (i.e. crystal) is the summation of scattering from all charged particles within the unit cell and is mathematically described by the following structure factor equation.

$$\mathbf{F}_{hkl} = \sum_{j=1}^{atoms} f_{s,j}^0 \cdot \exp(2\pi i(ha^* + kb^* + lc^*) \cdot r_j)$$

In the structure factor equation $f^0$ is the scattering factor, $s$ is the direction of the scattered wave relative to the incident beam $j$ is the type of scatterer and $r$ is the position of the scatterer The equation gives the total scattering from a crystal along a given lattice plane and is therefore proportional to the sum of all scattering contributions from all electrons in the unit cell and therefore directly related to the intensity of the corresponding reflection (Figure 3.14). The structure factor equation is dependent on the types of atoms $j$, and position of the $hkl$ reflection in reciprocal space ($ha^* + kb^* + lc^*$). This is where the phase problem arises. **F** itself cannot be calculated from the information provided in the diffraction experiment. Only the scalar magnitude $|\boldsymbol{F}|$ can be calculated from the intensity ($I_{hkl} = |\boldsymbol{F}_{hkl}|^2$).

$$\mathbf{F} = |\boldsymbol{F}| \exp[+i\alpha(hkl)]$$

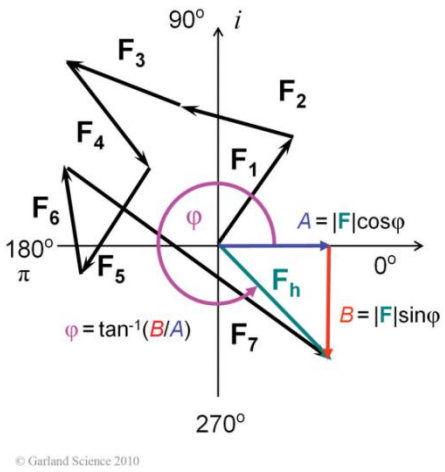



**Figure 3.14 Structure Factor**

Structure factor represented by an Argand diagram, having an imaginary vertical axis and a real horizontal axis. $F_{1\text{-}7}$ represents the contributions to the structure factor from seven individual atoms. The length of each black vector is determined by the atomic scattering factor. The angle to which the vectors contribute is determined by that atoms position between a set of reflecting planes. $F_h$ (green) is the resulting structure factor vector (From Biomolecular Crystallography by Bernhard Rupp, © 2012 Garland Science/Taylor & Francis LLC).

---

The relationship between the real crystal lattice and the observed reciprocal lattice can be interpreted using the Fourier transform. Because the Fourier transform allows us to move back and forth across the real and reciprocal space threshold we can reconstruct the real crystal lattice parameters from the reciprocal lattice, which is what is observed in the diffraction experiment. Likewise, if the **F** could be directly measured in the experiment one could immediately reconstruct the electron density in the unit cell using the Fourier transform.

$$\rho(xyz) = \frac{1}{\mathrm{V}} \iiint_{hkl} \mathbf{F}(hkl) \exp[-2\pi i(hx + ky + lz)] dhdkdl$$

Due to Bragg's Law, diffraction data is collected as numerous discrete spots. Each spot contains information about the global electron density along the particular scattering vector F(*hkl)*. Because all reciprocal lattice points are used to calculate the electron density ($\rho$) in the entire unit cell the above equation takes the form of a Fourier summation.

$$\rho(xyz) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}|F(hkl)|\exp[-2\pi i(hx+ky+lz)+i\alpha(hkl)]$$

For every *hkl* reflection the associated wave has two parts, an amplitude $|F|_{hkl}$, and a phase, $\alpha_{hkl}$. It is therefore possible to calculate the structure factor from the amplitudes measured in the diffraction experiment. However to complete the Fourier summation the phase must be also known. The phases for protein crystals cannot be directly measured and must be experimentally determined using indirect methods. This major hurdle is therefore known as the 'phase problem' in X-ray crystallography.

### 3.2.4 Anisotropic diffraction

In X-ray crystallography anisotropic diffraction is observed when the diffraction strength differs with cell direction. This can occur when the number and strength of the crystal contacts is less in one direction than the other. This gives rise to a higher crystal order in a particular orientation than the others (Figure 3.15).

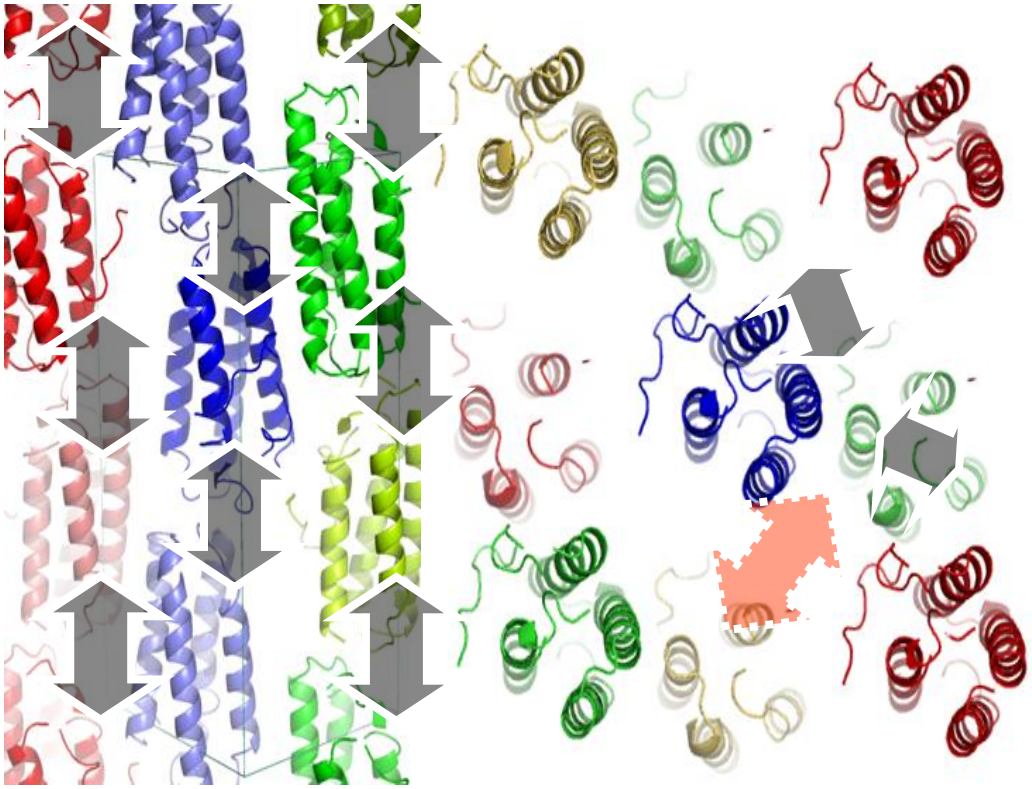


**Figure 3.15 Anisotropy**

Arrows represent the translation vectors between molecules. Left panel indicates all translation vectors are consistent along the vertical axis whereas the translation vectors in the horizontal plane (right panel) are not consistent, leading to anisotropy (with permission, Michael Sawaya, ACA 2006).

---

There are two major problems with anisotropic data. The first problem is that it can be difficult to decide where to make a resolution cutoff when processing the data. Ideally one would choose ellipsoidal resolution shell boundaries in which $I/\sigma$ and $R_{sym}$ will be comparable. Unfortunately data processing software does not readily divide data into ellipsoidal shells but nearly exclusively spherical ones. Anisotropic data affects resolution, therefore using a spherical shell on anisotropic data forces the researcher to either discard high resolution data containing structural detail or include poorly measured data which will compromise data processing statistics such as $I/\sigma$ and $R_{meas}$.

The second major problem arises during refinement because of the discrepancy between the observed structure factor, $F_{obs}$ and the calculated structure factor, $F_{calc}$ after a round of refinement. This is because $F_{calc}$ does not have any directional dependence unless explicitly designed into the refinement program. This is the reason that an anisotropic scale factor must be applied to the data before refinement in order for $F_{obs}$ and $F_{calc}$ to be comparable.

### 3.2.5 Twinning and Pseudosymmetry

Twinning occurs when a crystal consists of multiple domains which are mutually reoriented according to a specific transformation that does not belong to the symmetry operators of the crystal point group but is related in some way by the crystal lattice (Koch 1992). For proteins, this can result in the crystal growing with multiple unit cell orientations that are strictly described by a proper rotational operator. Merohedral twinning results when two or more non-equivalent orientations of the unit cell are permitted within the lattice. The condition where this can arise is when the symmetry of the crystal lattice is higher than the symmetry of the crystal point group. This phenomenon is permitted in tetragonal, trigonal, hexagonal and cubic crystal systems. Holohedry will be present in the presence of twinning. One must remember that the true symmetry of a crystal is defined by its content and not the geometry of the lattice. If the crystal class is a subgroup of lattice symmetry the unit cell can be rotated by a symmetry operator which obeys the crystal lattice but is not part of the crystal class. Diffraction patterns from this type of crystal will have perfectly overlapping reflections (identical *hkl* positions). The observed intensity of reflections will be the sum of different intensities from multiple non-equivalently oriented unit cells known as twin domains. Alpha (α) represents the ratio of one twin domain compared to another where the observed intensity is equal to: $\alpha \times I(hkl) + (1-\alpha) \times I(h'k'l')$ where *h'k'l'* are twin related indices.

A perfectly hemihedraly twinned crystal is composed of equal volumes of each twin domain. In this case detwinning is not possible. One can refine against twinned data simulating the presence of the two domains (Figure 3.14). When α is less than 0.5, detwinning can be attempted if the twin operator and twin fractions are known. Pseudomerohedral twinning occurs when serendipitous unit cell parameters emulate those of a higher symmetry system. For example if a

monoclinic cell that has a β angle very close to 90° may appear to be orthorhombic. With pseudomerohedral twinning the overlap of the twin lattices may not be perfect and broadening and splitting of reflections may be observed, particularly at higher resolution (Dauter 2003).

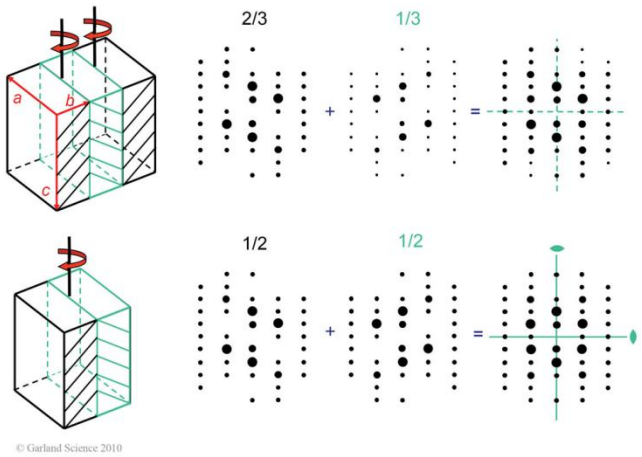



**Figure 3.16 Effects of Twinning on Diffraction**

The effects of twinning on the diffraction pattern from a crystal with 1/3, 2/3 twin fraction (top) and a 1/2, 1/2 twin fraction (bottom). Notice how perfect twinning (1/2, 1/2) leads to the addition of apparent 22 symmetry (From Biomolecular Crystallography by Bernhard Rupp, © 2012 Garland Science/Taylor & Francis LLC.).

---

Merohedral twinning cannot be detected by visually inspecting the diffraction images. Twinning can often be detected by analyzing the intensity distribution of a crystallographic data set. Theoretical intensity statistics for twinned crystals was pioneered by Stanley and later expanded on by Rees (Stanley 1955; Stanley 1972; Rees 1980; Rees 1982).

In general detection of twinning using intensity based statistics becomes more difficult when the twinning axis coincides with a translational non-crystallographic symmetry axis (NCS) also known as pseudo-symmetry. NCS gives rise to pseudo-symmetry when the NCS rotation axis is parallel to the crystallographic axis (pseudo-rotational symmetry) and/or the NCS translations are integer fractions of the unit cell vectors (pseudo-translational symmetry). Pseudosymmetry can cause higher apparent symmetry and abnormal intensity distributions.

There are known distributions for untwined and perfectly twinned data when the following twinning tests are applied.

#### 3.2.5.1 Cumulative Intensity Distribution of Z (N)Z

$Z = I/\langle I \rangle$: where $I$ is equal to intensity and (<I>) denotes the average intensity. (N)$Z$ for acentric reflections versus $Z$ yields a sigmoidal shape for twinned data.

#### 3.2.5.2 Perfect Twin Test

For the perfect twin test $\langle I^2 \rangle / \langle I \rangle^2$, where $I$ is equal to intensity and (<I>) denotes the average intensity, untwinned data typically has a value of 2 whereas perfectly twinned data has a value of 1.5.

#### 3.2.5.3 Partial Twinning Test

Partial Twinning Test: $H = |I\text{h} - I\text{h}'| / (I\text{h} + I\text{h}')$ where $H$ is a variable that is calculated by dividing the difference of two twin related reflections (h and h') by the sum of those same reflections ($I$ is equal to the reflections' measured intensities). $H$ can be calculated to give an indication of the twin fraction, denoted α, which can range from 0 to 0.5 in merohedrally-twinned crystals. This statistic requires knowledge of the twin operator so that the $I$ twin related reflections h and h' can be assigned (Yeates 1997,1988).

#### 3.2.5.4 *L* Test

The L test is more robust to anisotropy and pseudo-centering because it uses local pairs of intensities instead of using data binned in resolution shells. $L = (I_1 - I_2) / (I_1 + I_2)$ where $I_1$ and $I_2$ are not related by twinning or crystal symmetry. They are also near each other in reciprocal space. (Padilla and Yeates 2003).

### 3.2.6 Phasing

As mentioned previously the crystallographic phase problem arises from the inability to measure the relative phase of scattered waves. In the diffraction experiment we measure the intensity of a given reflection which relates to the square of the structure factor amplitude. Looking back at the Fourier summation equation, the exp[$i\alpha(hkl)$] term must be supplied by information from a separate phasing experiment. A few of the most common methods will be described below in

some detail. As a consequence a large amount of effort has been devoted to developing these phasing methods throughout history. There are several methods that have been developed to indirectly determine starting phases from which an initial electron density map can be generated.

#### 3.2.6.1 Single Wavelength Anomalous Dispersion (SAD)

Resonant scattering or anomalous scattering breaks down Friedel's Law which states that symmetry related reflections will have equal intensities meaning that symmetry related reflections are no longer equal. This happens when the incident photon energy is sufficient to knock out an inner shell electron creating a core hole. As a higher shell electron drops down to fill this core hole it emits a fluorescence X-ray (f") which has a phase shift of 90° relative to f' (non-anomalous scattering). This results in Bijvoet pairs which are Freidel pairs where the $\mathbf{F}^+$ and $\mathbf{F}^-$ are no longer equal.

In Figure 3.17 the anomalous scattering vector breaks the symmetry between $\mathbf{F}_A$ and $\mathbf{F}^-_A$ (second panel). The normal scattering contributions (green) are mirrored across the real axis while the anomalous contribution (purple) becomes mirrored across the imaginary axis.

A Multi-wavelength Anomalous Dispersion (MAD) experiment allows detection of dispersive differences between data collected at different wavelengths. This combined with the anomalous signal can increase the success of phase determination. Most commonly, data are collected at energies that correspond to the peak, inflection and remote features of X-ray fluorescence spectra collected near the absorption edge of the anomalous scattering atom. The most commonly used atom in MAD experiments presently is Selenium. It is used most often because it can be reliably incorporated with high efficiency.

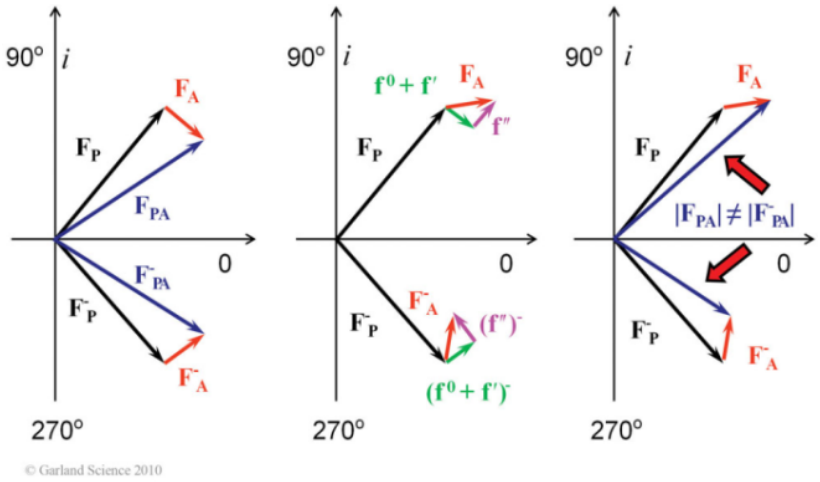



**Figure 3.17 Anomalous Differences**

The vector $\mathbf{F}_p$ represents the partial sum of all the non-anomalous scattering atoms. The vectors $\mathbf{F}_A$ and $\mathbf{F}^-_A$ represent the contributions from anomalous scattering atoms. The real component of the anomalous scattering atom is represented in green while the imaginary component is represented in purple (From Biomolecular Crystallography by Bernhard Rupp, © 2012 Garland Science/Taylor & Francis LLC.).

---

#### 3.2.6.2 Molecular Replacement

Molecular replacement (MR) utilizes the fact that one can make estimates of the phases using previously determined structures that well represents the unknown structure. Generally a sequence identity of 25 % or greater is required. Making a good guess of the position of the unknown molecule within the unit cell involves searching the unit cell volume. The search must be done in six dimensions; three rotations α, β and γ, and three translations x, y and z. In general the rotation parameters are determined first and the translation parameters second. The rotation search aims to identify the specific orientation of the molecules with respect to one another and the translations aim to place the model accurately into the real unit cell. The most popular method of conducting the rotation search is the application of the Patterson function.

$$\mathrm{P(uvw)} = \frac{1}{\mathrm{V}} \sum_{\mathrm{hkl}} |F(\mathrm{hkl})|^2 \cos[2\pi(\mathrm{hu} + \mathrm{kv} + \mathrm{lw})]$$

The Patterson function is the summation of the squared structure factors with the phase component of the equation removed, or rather, all phase angles set to zero. In the Patterson equation the *u, v* and *w* terms denote interatomic vectors. If these interatomic vectors are short they represent atoms within the same molecule. If these vectors are longer it may indicate these atoms belong to two different molecules. The shorter vectors, or self-Patterson vectors, are used to determine the rotation. The correct translations are found by searching x, y and z in the unknown unit cell using the probe molecule in a few predetermined rotational orientations and finding the position having smallest R-factor.

The MR search starts with the rotation search by placing the model inside an arbitrary unit cell at the origin, and calculating the Patterson map using the Patterson function. Then a Patterson map is calculated using experimental data. Then the two maps are rotated with respect to each other and the best fit is found. Once there is a strong correlation in the rotation function the translation function is initiated. This will determine the molecule's position within the real unit cell taking into account correct space group symmetry.

If the search is successful, the calculated phases can be combined with the observed intensities and the first electron density map can be calculated. One must not forget that phases dominate the electron density calculation which results in so called model bias which can be difficult to overcome especially when high resolution data is lacking. Two techniques can be applied in attempt to remove model bias. The first is simulated annealing (Brunger 1991) and the second is Iterative-build OMIT maps: map improvement by iterative model building and refinement (Terwilliger *et al.* 2008).

# 4.0 MATERIALS AND METHODS

**Table 4.1 Reagents, Supplies and Equipment**

| Item | Supplier |
|---|---|
| Ammonium Sulfate | Sigma Aldrich |
| Bis(2-hydroxyethyl)-amino-tris(hydroxymethyl)-methane (Bis-tris) | Fluka |
| Glycerol | Sigma Aldrich |
| Imidazole | Alfa Aesar |
| Isopropyl-β-D-thiogalactopyranoside (IPTG) | Sigma Aldrich |
| Tris (hydroxyl-methyl)-aminomethane hydrochloride (Tris-HCl) | Fluka |
| Tris (2-carboxyethyl)phosphine (TCEP) | Sigma Aldrich |
| Se-Met Kit | Medicilon |
| Sodium Chloride (NaCl) | Fluka |
| Sodium Malonate | Fluka |

## 4.1 Cloning

Plasmid DNA was obtained using the alkaline lysis method as described elsewhere (Sambrook *et. al.,* 1989). Recombinant DNA methods followed those described by Sambrook *et al.* (1989) including restriction endonuclease digestion, T4 DNA ligation and agarose gel electrophoresis.

The full open reading frame from amino acid 63-290 of ExeC had been previously cloned into the *Nco*I site of pVAC30 vector within the pRJ155.1 clone (Vivian Ast's Thesis, College of Microbiology and Immunology, University of Saskatchewan). This vector consisted of an 885 bp *Exe*C fragment ligated into the *Bam*HI site of the pET30a vector to produce construct later

referred to in this text as ExeC-ori (for original) (Figure 4.2). This vector served as the template DNA for further construct generation. The primers were designed to select for two different fragments of ExeC: amino acids 83-290 (1) containing both periplasmic domains, HR and PDZ and amino acids 83-181 (2) containing the HR domain. Vector pET-30a would also be used to produce the new C-terminally tagged HR and HR-PDZ constructs by ligation into the *Nde*I and *Xho*I sites to produce a more compact construct with a C-terminal tag (Figure 4.1). The primers were ordered from *Integrated DNA Technologies* and used to generate the desired gene product of ExeC by PCR amplification (Table 4.2).

**Table 4.2: List of Primers Used in Construct Design List of primers used in construct design**

Table of designed constructs, necessary primers, the incorporated restriction sites (bold) and vectors they were intended to be used with.

| Construct | Primers (Restriction in bold) | Restriction site | Plasmid | proteomics |
|---|---|---|---|---|
| HR-PDZ-C-tag (long) | 5'GAAG**CATATG**GCCCAGGCGGCCGATGC3'<br>5'CAAA**CTCGAG**TTCTGACAAGCCGACATAAACGTCGTA3' | *Nde*I<br>*Xho*I | pET-30a(+) | aa # 205<br>MW: 22152.7 Da<br>pI: 5.13 |
| HR-C-tag (short) | 5'GAAG**CATATG**GCCCAGGCGGCCGATGC3'<br>5'CCGG**CTCGAG**CAACGGCTTGCCGTACTCTTC3' | *Nde*<br>*Xho*I | pET-30a(+) | aa # 96<br>MW: 10347.5 Da<br>pI: 5.69 |

PCR amplification: The template DNA (1 µL) was used in a 50 µL reaction mix containing 0.5 µL *pfu* polymerase and 5 µL of 10 X *pfu* buffer, 80 µM dNTPs and the appropriate primer at a final concentration of 4 µM. The long fragments were successfully amplified using 30 cycles of denaturation (95°C, 1 min), annealing (55°C, 1 min) and extension (72°C, 2 min). The short fragments were successfully amplified by 30 cycles of denaturation (95°C, 1 min), annealing (60°C, 1 min) and extension (72°C, 1 min). The amplified fragments were separated and visualized on a 0.8% agarose gel stained with ethidium bromide under ultraviolet light. Once the

expected size of the amplicon was confirmed by visualization the remainder of the PCR reaction was purified from agarose using the *Geneclean II Kit* (*Qiagen gel extraction kit*).

*Xho*I and *Nde*I were the restriction sites designed into the primers for carboxy-terminal his-tagged constructs. Both the amplified fragment and the expression vector were separately digested with both restriction enzymes. Both the digested vector and digested DNA fragments were gel purified before ligation by T4 DNA ligase into pET-30a-C(+) (Novagen) (Figure 4.2). Ligations were performed in a 4°C water bath overnight. Primers were designed to introduce the desired restriction sites flanking the ExeC fragment of interest. The PCR product was isolated by running the sample on a 0.8% agarose gel followed by excision and purification of the fragment using a Qiagen gel extraction kit. Once the PCR fragment was cleaned with Min Elute Reaction Clean Up Kit from Qiagen, the isolated fragment was digested with the appropriate restriction enzymes prior to ligation.

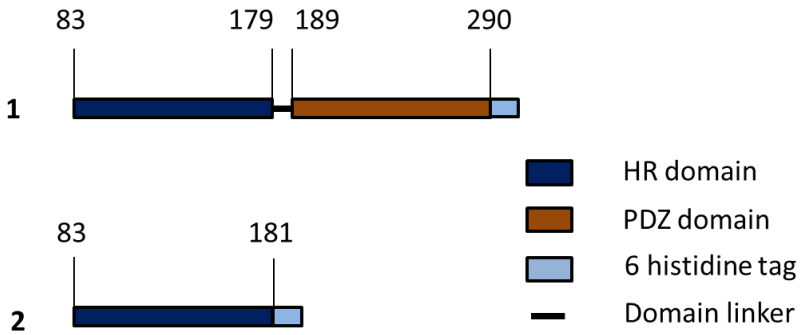



**Figure 4.1 Schematic of Newly Designed Constructs**

The schematic of both constructs. 1: HR-PDZ-C-tag and 2: HR-C-tag. These two constructs contain periplasmic portions of ExeC. The intent was to create a more succinct construct that would be more stable than ExeC-ori and more amenable to crystallization.

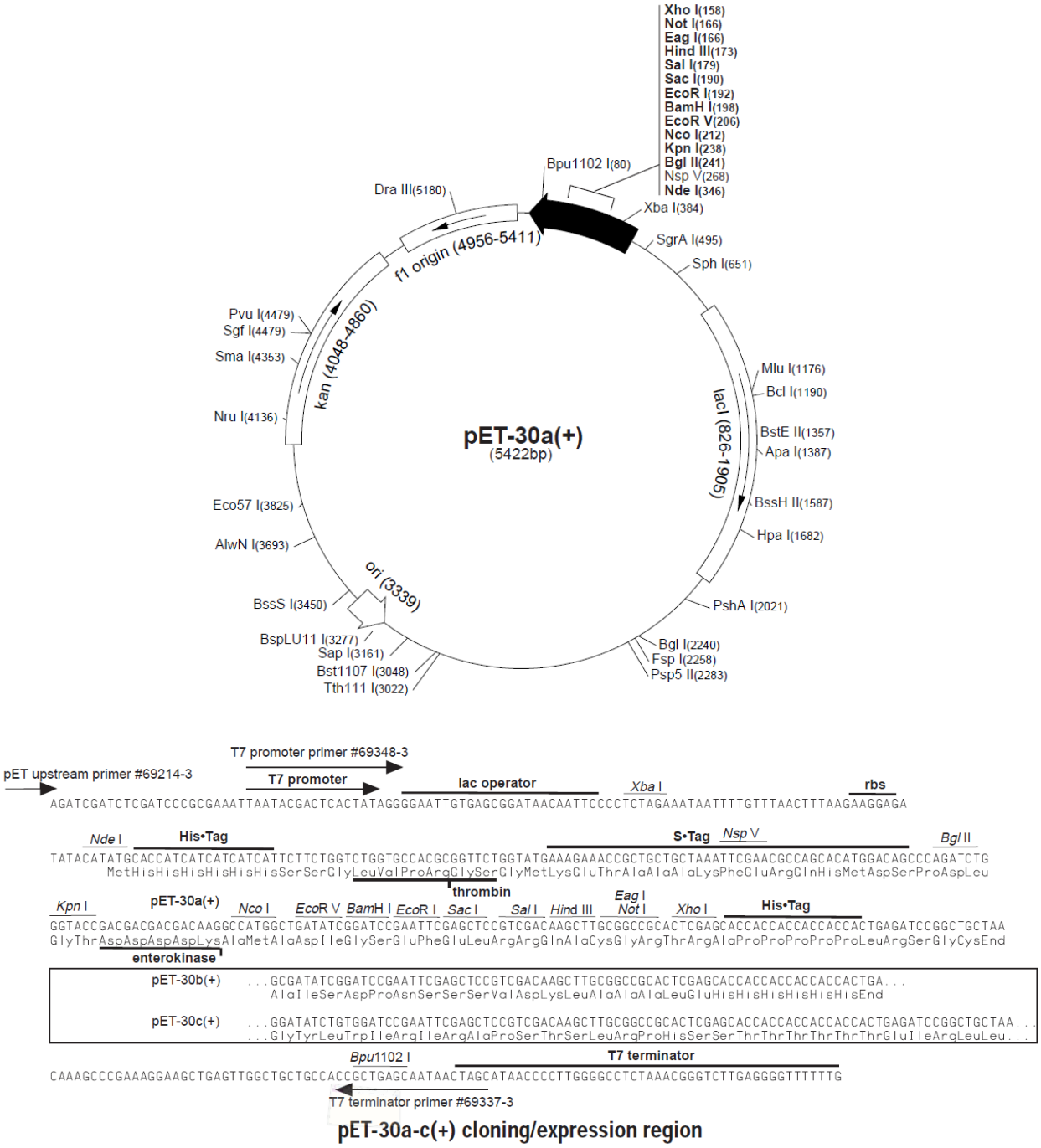



**Figure 4.2 pET-30a vector diagram**

The pET-30a vector diagram (top) along with the multiple cloning sites (bottom) was used to produce the three protein construct for this study (*Novagen*).

Transfection of the ligated plasmid was accomplished by electroporation into XL blue competent cells at 1.5 Volts time constant set to 4.4. Cells were immediately transferred to 450 ml of recovery media. Transformants were selected by plating them on an agar plate doped with 50 mg/L *Kanamycin*. PCR fragments were validated by DNA sequencing. Qiagen Plasmid Midi Kit was used to prepare large-scale amounts of plasmid DNA from XL1 blue cells and was submitted to the *Plant Biotechnology Institute National Research Council* for Sequencing and was found to have no errors.

### 4.2 Sodium-Dodecyl-Sulphate Poly-Acrylamide Gel Electrophoresis (SDS-PAGE)

SDS PAGE was run in a Mini-PROTEAN® 3 Mini vertical electrophoresis system (*Biorad*) at 150 volts. A discontinuous gel system was used, the stacking gel was 4 % polyacrylamide gel and the resolving gel was 12 %. Samples were mixed 5:1 with 5 X loading buffer containing 200 mM Tris-Cl (pH 6.8), 10% w/v SDS, 0.05 % w/v bromophenol blue, 20% glycerol and 10 mM β-ME. Snap-top micro-centrifuge tubes containing the samples were then placed in a boiling water bath for 10 min before being loaded in the gel. After the dye front was less than one centimeter from the bottom of the gel the voltage was turned off, the gel was removed from the glass cassette and rinsed thoroughly with distilled water before being transferred into Coomassie Brilliant Blue. Once in the stain gels were microwaved for one minute at full power then placed on an orbital shaker. After five minutes the stain and gel were heated gain for 30 seconds in the microwave and again placed on the orbital shaker. After staining is complete the gel was transferred to distilled water for de-staining. The molecular weight marker used for this study is the kaleidoscope protein ladder (*BioRad).* The size of each band from the bottom up is as follows; (kDa) 10, 15, 20, 25, 37, 50, 75, 100, 150 and 250.

### 4.3 Over Expression

For protein expression the desired vectors were electroporated into XL1 blue competent bacterial cells and the transformants were selected for as per the above procedure. Single colonies were used to inoculate a 5 ml of 2 x YT culture media and incubated overnight in a 37°C shaking water bath (Sambrook *et. al.,* 1989). 1 mL of the saturated overnight culture was used to

inoculate 800 mL cultures. Induction of cell cultures was done at an optical density of 0.7 at 600 nm ($OD_{600}$) by adding IPTG at a final concentration of 1 mM. After 4 hr. of shaking in a 37°C incubator, cells were harvested by centrifugation for 10 min at 5 000 x g.

#### 4.3.1 Seleno-Methionine Derivitization

The same BL21 (DE3) cells were used in a kit purchased from *Medicilon* designed for this purpose which follows the method first described by Doublie (Doublie 1997). The kit follows the method first published by (Doublie 1997) and is based on the use of minimal media and methionine synthesis inhibitory drugs.

### 4.4 Protein Purification

#### 4.4.1 Immobilized Metal (Ni) Affinity Chromatography (IMAC)

Frozen cell pellets were thawed on ice before being re-suspended in 20 mL of binding buffer (50 mM Tris-HCl pH 8.0, 50 mM $MgCl_2$, 250 mM NaCl, 10 mM Imidazole and 0.5 mM PMSF) containing 1 tablet of *complete protease inhibitor EDTA-free* (Roche) as well as 0.2 mg of RNase and 0.4 mg of DNase. Re-suspended cells were then passed through a chilled French press two times followed by two centrifugations at 20 000 x *g* for 30 min.

The supernatant was then loaded on a fully equilibrated 5 mL Ni IMAC column and washed with 400 mL of binding buffer. ExeC was then eluted with an elution buffer gradient consisting of binding buffer with an additional 490 mM Imidazole. The gradient went from 0%-100% elution buffer over 100 mL. Although the protein eluted from the column below 50%, the gradient was allowed to continue for column cleaning purposes.

IMAC involves a solid phase chromatography column and an engineered protein. The efficacy of this method results from the unique properties of a polyhistidine motif engineered onto the terminus of the protein. The cation is immobilized on the resin in the column and the polyhistidine tag binds to the cation via chelation. All non-binding proteins will wash free of the column. The binding interaction is strongest at pH 8 and can be disrupted by changing the pH or,

more commonly, by the addition of imidazole which at a high enough concentration will compete with the tagged protein binding.

### 4.4.2 Gel Filtration Chromatography

Gel filtration, also known as size exclusion chromatography acts as a molecular sieve to separate proteins according to their globular size. Gel filtration was performed using a *superdex 75 16/60 column* (GE healthcare). The sample was concentrated to a volume of 1 mL using an *Amicon Ultrafiltration* device before injection to the column fully equilibrated with buffer containing 20 mM Tris-HCl pH 7.0 and 150 mM NaCl (1 mM TCEP was used with Se-Met ExeC). The flow rate was set at 0.5 mL/min while fractions were collected in 2 mL aliquots.

### 4.4.3 Ion Exchange Chromatography

Anion exchange chromatography separates biomolecules according to the net charge on the surface of the molecule. The solid phase will have positive charge for anion exchange and negative charge for cation exchange. One must optimize the pH so that the molecule of choice has the appropriate charge. For example, ExeC-Long has a theoretical isoelectric point of pH 5.13 so a pH above this will cause it to carry a net negative charge. This charge causes it to bind to the cationic resin within the column. The addition of high ionic strength buffer will disrupt this interaction and elute the bound protein.

Anion exchange was done using both the Mono-Q and Q-FF columns (GE Healthcare). Columns were fully equilibrated with binding buffer containing 20 mM Bis-tris pH 6.2 and eluted with an increasing gradient of NaCl.

## 4.5 Dynamic Light Scattering

Dynamic Light Scattering (DLS) is an important tool in assessing purified proteins before crystallization trials begin. Homogeneity is a prerequisite for crystallization and DLS can measure the hydrodynamic radius and therefore size distribution of very small particles. Ideally, the protein intended for crystallization will have a monodispersed profile. A polydispersed profile indicates that there is more than one conformational species present which may consist of

a mix of monomers and multimers. DLS analysis on ExeC long was performed at one mg/mL in 100 mM NaCl and 20 mM Tris pH 7.5 using the DynaPro-MS800 at the *Saskatchewan Structural Science Centre.*

## 4.6 Mass Spectrometry

Initially, mass spectroscopy (MS) was used to determine the more stable fragment from the original construct which guided the design of subsequent constructs. The sample was sent to *WEMB Biochem Inc*. as slices from poly-acrylamide gels, (Figure 4.2) these were then analyzed by MS sequencing.

The mass spectrometer at the *Saskatchewan Structural Science Centre* is an API Qstar XL pulsar hybrid lc/ms/ms and it was used to analyse purified protein samples from FPLC. The Qstar system utilizes quadrupole technology along with a time of flight mass analyzer. Protein solutions of ExeC were injected into the ionspray ioniser via the integrated syringe pump. Result summary are shown in the results and discussion section.

## 4.7 Crystallization

### 4.7.1 Protein Preparation

Peak fractions collected in the final step of purification were pooled together and concentrated using an *Amicon Ultrafiltration* device centrifuged at 20 000 x g. Once all of the pooled fractions were able to fit into the 15 mL device it was centrifuged until the protein pool was concentrated to a volume of 1 mL, then 15 ml of crystallization buffer was added to the device resulting in a 15 X dilution of the original purification buffer. Repeating this procedure several times results in near complete buffer exchange.

### 4.7.2 Vapor Diffusion Set Up for Crystallization of ExeC

Typically Vapour diffusion crystallization trials were set up in 24 well *Nextal* (now Qiagen) 'EasyXtal Tool' using the hanging drop method. Greaseless screw caps were blown off with a puff of compressed gas using 'Super Duster' made by *MG Chemicals*. The cap was then inverted and 1 µL of protein solution was mixed with 1 µL of well solution prior to sealing the well.

### 4.7.3 Screening

Screening for crystallization conditions was done using commercial kits produced by Nextal ® and Qiagen ®. Each kit consisted of 96 conditions in 10 ml tubes. Plates were inspected one day after set up and periodically after that until the wells dried up. Any condition that produced an interesting result was typically used to set up a multifactorial plate surrounding the initial condition. A table of screening conditions can be found in Appendix 1.

### 4.7.4 Cryoprotection of Crystals

Crystals were plunged into liquid nitrogen after being removed from the crystallization drop with a nylon loop under a dissecting microscope. For the ExeC crystals several different cryoprotectants were tried including PEG 400, sodium malonate, oil as well as the most common cryoprotectant, glycerol. Various strategies were used while cryoprotecting crystals. The most basic of these is picking the crystal from the crystallization drop and dipping it in another drop containing up to 20% of the cryoprotectant. The crystal was then removed from this drop and plunged in to liquid nitrogen. More complex methods included moving the crystal through a series of drops starting at low cryoprotectant concentration up to a final drop containing the highest concentration of cryoprotectant in a series of steps. One can do this process at different rates allowing the crystal to incubate for longer or shorter duration in each drop before moving to the next. When Oil was used a small amount was dispensed directly on top of the crystallization drop and the crystal was separated from the mother liquor and picked out of the oil near the edge of the drop.

## 4.8 Diffraction Data Collection and Processing

All X-ray diffraction data were collected at the *Canadian Macromolecular Crystallography Facility* (CMCF) at the *Canadian Light Source* (CLS) beamline O8ID-1 (CMCF-ID) equipped with the Oxford Instruments Cryojet. Data presented here was processing using *XDS* (Kabsch 1993) *Autoprocess* and associated programs. Additional processing was carried out with *HKL2000* (Otwinowski and Minor 1997), and, in some cases, d*trek (Pflugrath 1999) (data not shown).

### 4.8.1 Anisotropic Correction

Anisotropic correction was performed as a subroutine in *Truncate* (*CCP4*) (Padilla and Yeates 2003) and by the online server at UCLA (http://www.doe-mbi.ucla.edu/~sawaya/anisoscale/).

### 4.8.2 Analysis of Diffraction Data for Crystal Twinning

*TRUNCATE* (*CCP4*) was the primary program used in this study to implement twinning tests. Programs such *SFCHECK*, *DETWIN*, *TRUNCATE* (*CCP4*) and *XTRIAGE* (*PHENIX*) were also used to test diffraction data for twinning (data not shown).

## 4.9 Phasing

### 4.9.1 Molecular Replacement

Two search models (PDBs; 2I6V and 2I4S) were used for MR probes. They both have greater than 40% identity with the PDZ domain of ExeC. *PHASER* (*CCP4*) was primarily used through the *CCP4*i interface. *MRBUMP* (*CCP4*) and *BALBES* (web server from York University: http://www.ysbl.york.ac.uk/YSBLPrograms/ index.jsp) were also used. *BALBES* only uses *MOLREP*, whereas *MRBUMP* utilizes both *PHASER* and *MOLREP*. Models of varying lengths were created primarily with CHAINSAW (*CCP4*). Solvent content and resolution range were systematically varied in an attempt to find a valid solution.

### 4.9.2 Multi-wavelength Anomalous Dispersion (MAD)

Experiments for crystals containing ExeC derivatized with Seleno-methionine were similar to those used for native ExeC except that the diffraction experiment was conducted using an incident X-ray beam tuned to the K absorption edge of Se. XAS scans across the Se K-edge placed the peak energy at 0.9795 Å while the inflection point energy was 0.9797 Å. Several programs were used to try and exploit the anomalous signal present in SeMet data sets. *SOLVE/RESOLVE* and its predecessor *CNS* was used alongside the *SHELX* and *PHENIX* suite of programs.

# 5.0 RESULTS

## 5.1 Cloning of ExeC Constructs

Both of the constructs HR-PDZ-C-tag and HR-C-tag (Figure 4.1) were successfully produced using the method described in section 4.2.

## 5.2 Over-expression of ExeC

Over-expression was successful for all three constructs. SDS-PAGE analysis on cleared cell lysate showed a predominant band that was of the size expected for both HR-PDZ- C-tag and HR- C-tag constructs (Figure 5.1).

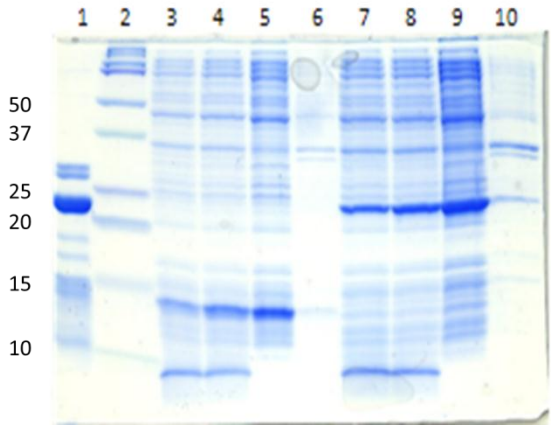



**Figure 5.1 Over-Expressed ExeC (SDS-PAGE)**

SDS-PAGE of whole cell lysates for clones HR-C-tag and HR-PDZ-C-tag. Lane 1: previously purified PDZ-ori incubated at 4° for several months. The boldest band is the naturally occurring 'stable' fragment that inspired the creation of the new constructs. Lane 2: MW marker from the bottom up (kDa) 10, 15, 20, 25, 37, 50, 75, 100, 150 and 250. Lanes 3, 4: whole cell lysates of cells expressing HR-C-tag construct 2 and 4 hr. After induction by addition of IPTG. Lanes 5, 6: soluble and insoluble fraction of harvested cells after lysis and centrifugation. Lanes 7, 8: Whole cell lysates of cells expressing HR-PDZ-C-tag construct 2 and 4 hr. after induction by addition of IPTG. Lanes 9, 10: soluble and insoluble fraction of harvested cells after lysis and centrifugation.

## 5.1 ExeC Instability

The newly-designed HR-PDZ- C-tag ExeC construct was also unstable. Storage at low temperatures and the use of a protease inhibitor cocktail did not prevent ExeC from degrading into smaller MW fragments. In fact the same band pattern was observed for both constructs that contained the two ExeC periplasmic domains. Using HR-PDZ- C-tag, which was significantly degraded, it was possible to determine which of the two predominant low MW fragments was the C-terminal (PDZ) portion and which was the N-terminal (HR) portion of the original construct. Making use of the carboxy terminal six histidine tag it was possible to use IMAC and SDS-PAGE to decipher which fragment contained the C-terminal tag and which did not. The presence of 100 mM NaCl hampered degradation, however all crystallization attempts in the presence of NaCl were unsuccessful. The linker between the HR and PDZ domains was susceptible to proteolytic cleavage (Figure 5.2 ).

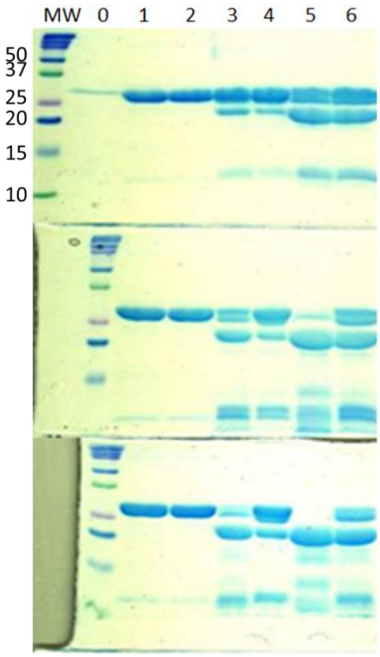



**Figure 5.2 ExeC Degradation (SDS-PAGE)**

The Right hand six lanes are three different ExeC-ori purification preps. The first, lane 1 & 2 was one day old lane 3 & 4 was 2 weeks old stored at 4°. Lane 5 & 6 two months old stored at 4°. The top gel represents the state of the different protein preps at the initiation of the experiment. The middle gel was prepared 20 hours later and the third gel was prepared at 44 hours after the initiation. Purification preps were ran as pairs. The left hand partner being incubated at room temperature and the right hand partner at 4°C. Lane MW) molecular weight marker. Lane 0) top panel only is not part of the study.

---

The degraded sample was applied to a mini spin Ni NTA column. The lowest weight band (~11 kDa) bound to the resin indicating it is the PDZ domain (Figure 5.3). This eventually allowed for the identification of the crystallisable fragment as the PDZ domain using crystallization on both low MW fragments. As a result, only the lowest MW fragment could be used to grow the familiar hexagonal crystals.

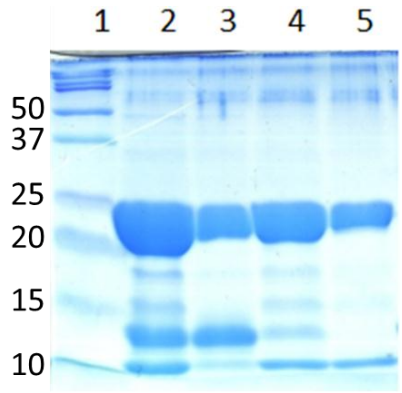



**Figure 5.3 Degraded ExeC Ni-NTA Binding Assay**

This gel indicates which of the two predominant low MW bands contains the C-terminal six-his tag. Lane 1: MW marker. Lane 2: Load solution. Lane 3: FT. Lane 4: $H_20$ wash. Lane 5: imidazole elution. One can immediately notice that the band at ~14 kDa does not bind the resin and comes out in the FT while the band at ~11 kDa binds to the resin.

---

## 5.2 Protein Purification of ExeC-HR-PDZ

The capture step of the purification process was Ni IMAC. This step needs to have high yield and is usually the first step in the purification procedure. The cleared lysate was transferred into a 50 mL falcon tube and diluted with loading buffer to reduce viscosity and ensure consistent loading parameters. One peak was observed eluting at ~25 % elution buffer, the elution peak typically has a volume of ~40 ml. SDS-PAGE analysis of the peak fraction (Figure 5.4) revealed that there was a large proportion of higher MW contaminants present. Examination of the gel suggested that the intensity of the high MW contaminants were directly proportional to the intensity of the desired 22.12 kDa target protein. To further support this hypothesis, purified samples which started out at 95 % purity (judged by visual inspection on SDS-PAGE gels) would eventually show both lower MW fragments as well as the high MW pair of bands after being stored at 4° C. The high MW bands have relative masses of approximately 67 and 75 kDa. What this may suggest is that ExeC might form a trimer that is recalcitrant to denaturation throughout SDS-PAGE analysis.

Gel filtration proved to be a valuable purification step as it was able to fractionate both higher and lower molecular weight proteins from the desired full length monomer. Both ExeC long and short were found to have different retention times depending on the amount of NaCl present in the gel filtration buffer. In the absence of any salts, ExeC ran as a multimer whereas the addition of 150 mM NaCl to the filtration buffer caused the elution peak to shift to that of the lower molecular weight (results not shown).

The *mono Q* (*GE Healthcare*) was a very powerful tool for polishing the ExeC extract after the capture step. It was shown that both higher and lower MW contaminants were successfully removed resulting in substantial amounts of highly pure ExeC. The higher MW contaminants were present after purification using the Ni column, but could be removed with ion exchange (Figure 5.5) or gel filtration (Figure 5.6). One advantage that the mono Q column could provide overt the gel filtration column was the ability to resolve proteins that are of similar MW. For instance, the smaller degraded fragments from aged ExeC preparations could be separated from one another.

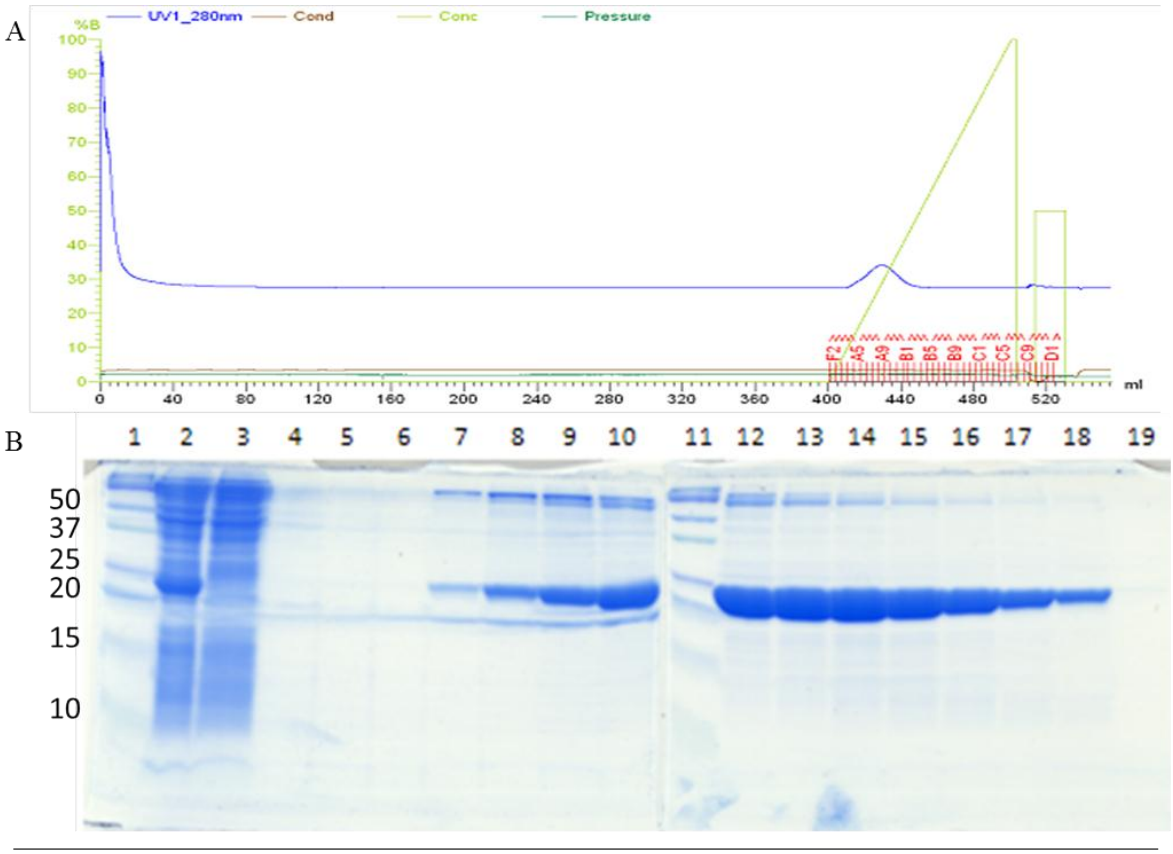



**Figure 5.4 Ni IMAC**

A: Chromatogram from a typical Ni IMAC purification. The blue trace represents absorbance of UV light at 280 nm (mAu). At 0 ml the wash step initiates pumping loading buffer through the column with the intention of washing out all proteins that are not specifically binding the column resin. After washing with 20 x the column volume, the elution gradient was initiated. This is represented by the green line and indicates the percent of elution buffer washing the column. The red dashed numbers represent the collected fractions. B: SDS-PAGE of fractions, Lane 1: MW marker. Lane 2: The cleared lysate in loading buffer just prior to being loaded on the column. Lane 3: the flow through before washing. Lane 4-6: flow through collected during the wash at 100, 200 and 400 ml respectfully. Lane 7-18: the eluted peak collected in fractions A4-b2. Lane 19-20: fractions B7 and B10.

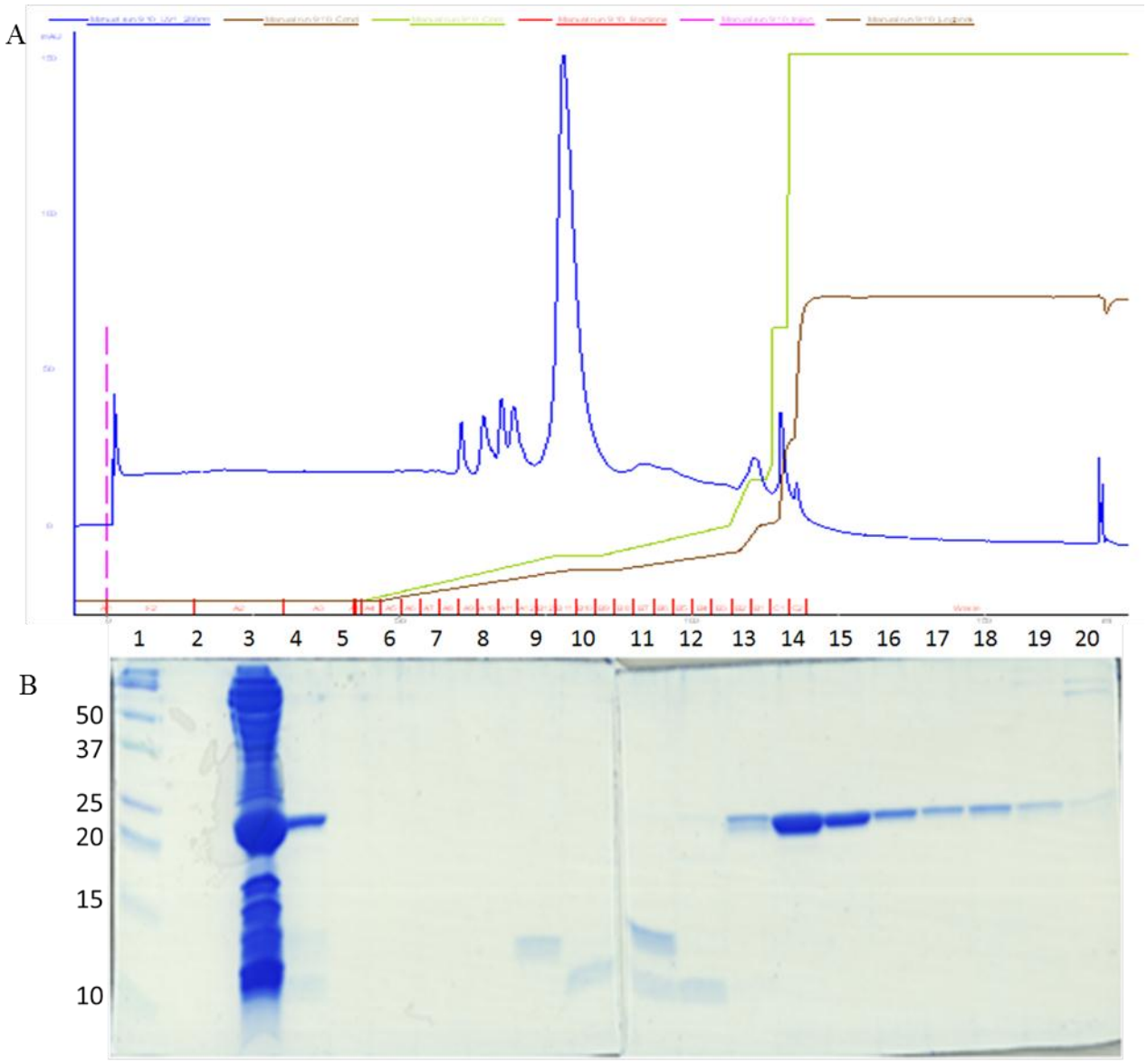



**Figure 5.5 Anion Exchange**

A: The blue trace represents absorbance of UV light at 280 nm (mAu). After washing with 20 x the column volume the elution gradient was initiated. This is represented by the green line and indicates the percent of elution buffer used to wash the column. The red dashed numbers represent the collected fractions. B: SDS-PAGE of fractions, Lane 1: MW marker. Lane 2: blank. Lane 3-4: load solution concentrated and diluted. Lane 5-20: peak fractions collected starting at the beginning of the elution gradient for 16 consecutive 3.33 ml fractions.

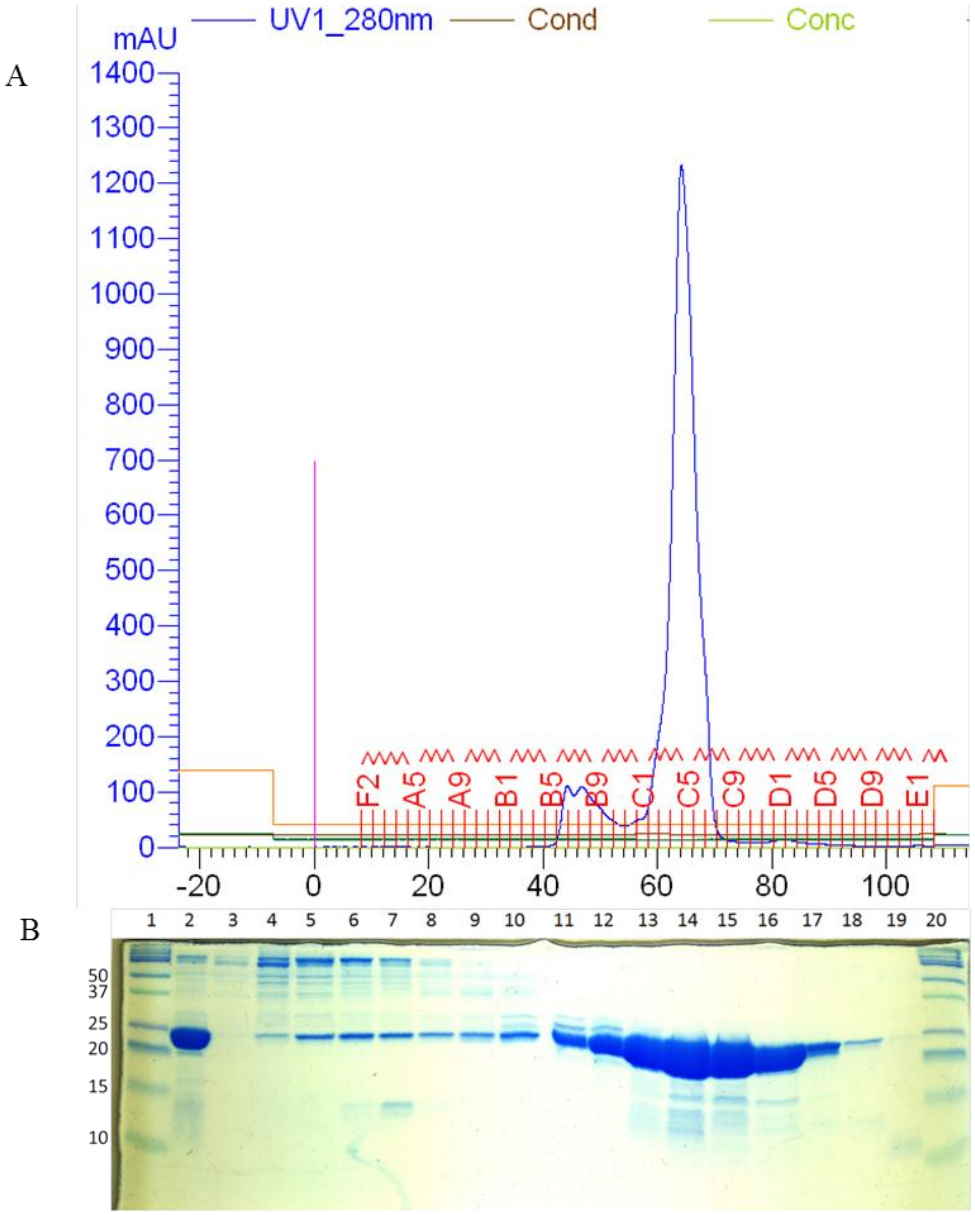



**Figure 5.6 Gel Filtration**

A: The blue trace represents absorbance of UV light at 280 nm (mAu). Gel filtration effectively removed larger proteins. Notice how the peak fractions already seemed to contain degradation contaminants by the time SDS-PAGE was run. B: SDS-PAGE of fractions, Lane 1: MW marker. Lane 2: load solution. Lane 3-18: fractions B5 – C8. Lane 19: fraction D1. Lane 20: MW marker. The blue line represents absorbance of UV light at 280 nm. The elution volume of the main peak is 60 mL as shown by the black numbers below the chromatogram.

## 5.3 Dynamic Light Scattering on ExeC-HR-PDZ

The results of dynamic light scattering showed the presence of one predominant peak. The secondary peak was 14% of the main peak measured by signal intensity. A 14 % polydispersity reading is regarded as a monodispersed solution. The measured hydrodynamic radius was that of a 48.66 kDa globular protein. Interestingly this data suggests that HR-PDZ- C-tag may exist as a dimer in solution.

## 5.4 Mass Spectroscopy of ExeC-HR-PDZ Fragments

In order to determine accurate molecular mass measurements for the degradation fragments, ExeC was analysed by mass spectroscopy. In addition, it was anticipated that incorporation of SeMet could be verified using this method (Figure 5.7).

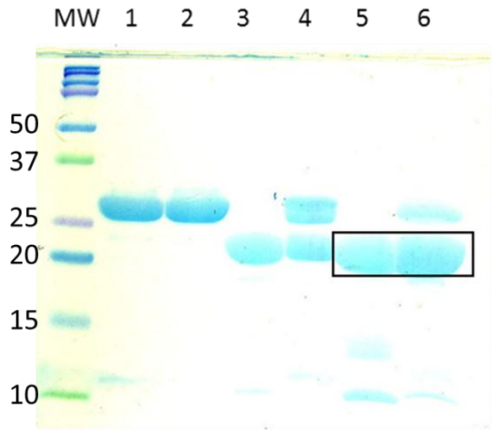



**Figure 5.7 MS Gel Fragment**

SDS-PAGE gel of 3 different samples at room temp and at 4 ºC. Lane 1: MW. Lane 2, 3: low protein concentration with NaCl, theses lanes represent the size of the complete construct. Lane 4, 5: High protein concentration; no NaCl. Degradation is occurring at room temperature more rapidly than at 4 ºC Lane 6, 7: high protein concentration with NaCl. The rate limiting step was interpreted as a more stable ExeC fragment. The black square indicates the excised bands used for mass spec.

The mass of the native PDZ domain determined in this experiment defines the crystallized fragment (Table 1.). The measured data for the full length native protein is 113 Da heavier than the theoretical mass of the protein. It is hard to say with certainty what this discrepancy means however it may indicate the presence of a ligand. The PDZ domain results could suggest that each molecule may only contain one Se atom, or that the cleavage of the two domains is not the same for the native and derivatized protein. With two missing variables; the amount of Se incorporation and the unknown amino acid composition of the derivatized PDZ domain, makes it very difficult to form solid conclusions regarding these results. To add to the uncertainty is the much higher than expected measured weight of the derivatized HR domain.

**Table 5.1: Summary of Mass Spectroscopy Results.**

The following table summarises the mass spec results performed at the SSSC. The mass in Daltons (Da) is reported along with the corresponding residue range (corresponding to figure 3.9) for the given value. Theoretical values for each fragment are based solely on MS results

| Fragment | Native | SeMet | Theoretical | |
|---|---|---|---|---|
| | | | Native | SeMet |
| Full length | 22269 Da | 22551 Da | 22156 Da | 22708 Da |
| PDZ | 11466 Da<br><br>Residues<br>193-290+6-his tag | 11485 Da<br><br>(dependent on SeMet incorporation) | 11466 Da | 11701 Da |
| HR | 10279 Da<br><br>Residues<br>93-188 | 10884 Da<br><br>Residues<br>93-191 | 10279 Da | 10373 Da |

## 5.5 Crystallization and Cryoprotection of ExeC-HR-PDZ Crystals

Crystals were originally found in Condition # 86 of the JCSG suite (*Joint Center for Structural Genomics*) approximately two months after the crystallization trial was initiated. A series of plates were subsequently set up around the condition; 1 M Ammonium sulphate, 0.1M bis-TRIS

pH 5.5 and 1% w/v PEG 3350 to further improve the crystallization conditions. Using a variety of protein preparations and two different temperatures, it was determined that there were two different conditions that produced crystal.

The first crystal form was reproduced with protein eluted from the gel filtration column in 20 mM Tris pH 8 and plated at concentrations between 10 and 20 mg/mL. Crystals grew in 1.0 - 1.25 M $NH_4SO_4$ and 100 mM Bis-Tris pH 5.5 first at room temperature and later at 4 °C. The addition of 1 % PEG 400 and 1 % PEG 3350 to the crystallization buffer also permitted crystal growth, however with worse diffraction quality. Additives such as malonate and citrate among other carboxylic acids seemed to induce too much nucleation leading to showers of crystals in the crystallization drops instead of useable single crystals. This hexagonal crystal form would produce several diffraction data sets of varying quality (Figure 5.8).

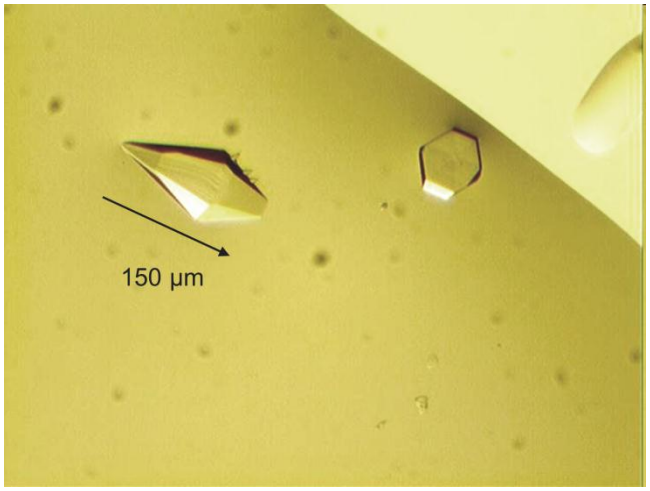



**Figure 5.8 Hexagonal bi-Pyramidal ExeC Crystals**

Hexagonal bi-pyramidal protein crystals of the PDZ domain of ExeC. The larges crystals grew to 150x50x50 µm in size. The crystals shown here have sharp edges which can be an indication of good crystal quality. Often crystals without sharp edges formed resembling a football, these crystals were found to diffract poorly.

---

The second crystal form was found in the following condition: 5% PEG 3350, 100 mM Bis-Tris pH 5.5 at 4°. These crystals diffracted to better than 3 Å in one direction but at 90° from this

orientation showed powder-like diffraction. SDS-PAGE analysis of these crystals clearly showed that they are composed of an 11-12 kDa fragment. A single crystal was also found in the cold room in a similar condition; 8% PEG 3350, 100 mM Bis-Tris pH 5.6, after a two year incubation. The morphology is different than the above mentioned crystals. This single crystal was found to be protein; however the diffraction was very weak.

Cryoprotection using glycerol was often successful at preventing ice rings in the diffraction pattern however streaking of spots in the *a*/*b* plane often occurred. Similar results can be reported for malonate and PEG 400 although in some instances these latter two compounds were believed to be responsible for unit cell doubling.

### 5.6 Diffraction and Data Processing of Hexagonal ExeC Crystals

Out of approximately 100 crystals screened at the Canadian Macromolecular Crystallography Facility (CMCF) beamline 08ID-1, 44 crystals were used to collect data sets. The quality of the data collected shows a number of problems. The diffraction is anisotropic and the crystals are readily damaged by radiation making it difficult to collect a complete data set. The most significant problem is that the correct space group cannot be determined with absolute certainty when the possibility of twinning and/or pseudo-symmetry is considered.

Data processed with *AutoProcess* consistently places the data in the $P6_222$ / $P6_422$ space group with unit cell dimensions a=b= 50 Å, c= 150Å, $\gamma$= 120°. When a self-rotation function is calculated from the diffraction data it appears that there is indeed 622 point-group symmetry or 6/mmm Laue group symmetry. To complicate matters these crystals were found to have a second unit cell with a doubled c axis length (a=b=50Å c=302Å, $\gamma$= 120°). As mentioned before cryoprotection is a believed to be a factor in the unit cell changing. Crystals from the same drop treated with different cryoprotectants would display large and small unit cell parameters. It should be noted that test images taken of crystals mounted at room temperature in a MicroRT (*Mitegen*) set up displayed the large unit cell.

Below diffraction data from three different crystals are shown as examples. Crystal Nat-Small is a native crystal and displaying the small unit cell. Crystal SeMet-Small is a selenomethionine-

derivatized crystal that also displays the small unit cell. Crystal Nat-Large is a native crystal however this crystal displays the large unit cell.

**Table 5.2: ExeC Data Collection Statistics of Crystal Nat-Small**

| Nat-Small | | | |
|---|---|---|---|
| Space Group | *P*3 | $P6_4$ | $P6_422$ |
| Wavelength (Å) | 0.97934 | 0.97934 | 0.97934 |
| *a* | 52.3636 | 52.3636 | 52.3636 |
| *b* | 52.3636 | 52.3636 | 52.3636 |
| *c* | 149.8046 | 149.8046 | 149.8046 |
| γ | 120 | 120 | 120 |
| Resolution (Å) | 12.97-2.90(2.98-2.90) | 12.97-2.90(2.98-2.90) | 12.97-2.90(2.98-2.90) |
| All Reflections | 25769 (1882) | 25862(1880) | 25945(1891) |
| Unique Reflections | 10032(739) | 5172(373) | 3091(212) |
| Multiplicity | 2.6 | 5.0 | 8.4 |
| Completeness (%) | 98.3 (99.9) | 99.5 (99.7) | 99.5 (99.5) |
| Mosaicity | 0.21 | 0.21 | 0.21 |
| I/Sigma(I) | 14.64(1.99) | 20.17(2.79) | 26.24 (4.11) |
| $R_{meas}$* | 4.9 (48.2) | 5.0(49.9) | 5.0 (39.1) |

These statistics are a result of processing diffraction images with *AutoProcess* at CMCF-ID with Friedels law set to True. *Rmeas (redundancy independent Rfactor (Diederichs & Karplus, 1997, *Nature Struct. Biol*. **4**, 269-275.)). Rmeas = Sum(h) [ Sqrt(n/(n-1)) Sum(j) [I(hj) - <I(hj)>] / Sum(hj).

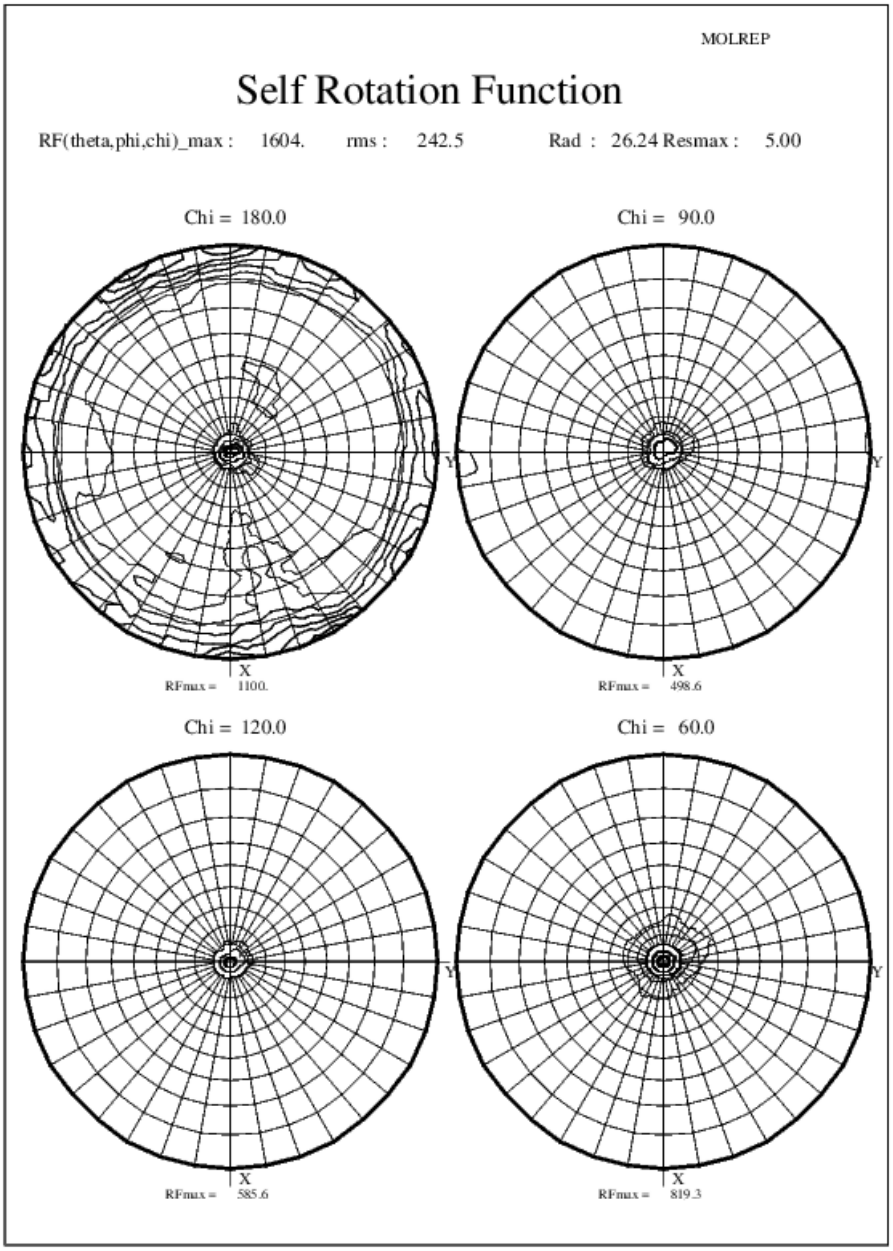



**Figure 5.9 Self Rotation Function of Data Set Nat-Small**

The data used in the self-rotation function was processed in space group *P*1 and between 13-5 Å resolution using *MOLREP*.

**Table 5.3: ExeC Data Collection Statistics of Selenomethionine Derivatized Crystal SeMet-Small**

| SeMet-Small | | | |
|---|---|---|---|
| Space Group | *P*3 | $P6_4$ | $P6_422$ |
| Wavelength (Å) | 0.97952 | 0.97952 | 0.97952 |
| *a* | 49.6750 | 49.6750 | 49.6750 |
| *b* | 49.6750 | 49.6750 | 49.6750 |
| *c* | 151.5674 | 151.5674 | 151.5674 |
| γ | 120 | 120 | 120 |
| Resolution (Å) | 13.42-3.00(3.08-3.00) | 13.42-3.00(3.08-3.00) | 13.42-3.00(3.08-3.00) |
| All Reflections | 47107 (3511) | 46946(3512) | 46883(3512) |
| Unique Reflections | 8415(646) | 4261(326) | 2570 (190) |
| Multiplicity | 5.6 | 11.0 | 18.2 |
| Completeness (%) | 100 (100.5) | 100 (100.6) | 100 (100.5) |
| Mosaicity | 0.195 | 0.195 | 0.195 |
| I/Sigma(I) | 14.08(2.21) | 20.05(3.11) | 25.81 (4.10) |
| $R_{meas}$* | 8.4 (79.7) | 8.3(79.9) | 8.2 (79.5) |

These statistics are a result of processing diffraction images with *AutoProcess* at CMCF-ID with Friedels law set to True. *Rmeas (redundancy independent Rfactor (Diederichs & Karplus, 1997, *Nature Struct. Biol*. **4**, 269-275.)). Rmeas = Sum(h) [ Sqrt(n/(n-1)) Sum(j) [I(hj) - <I(hj)>] / Sum(hj).

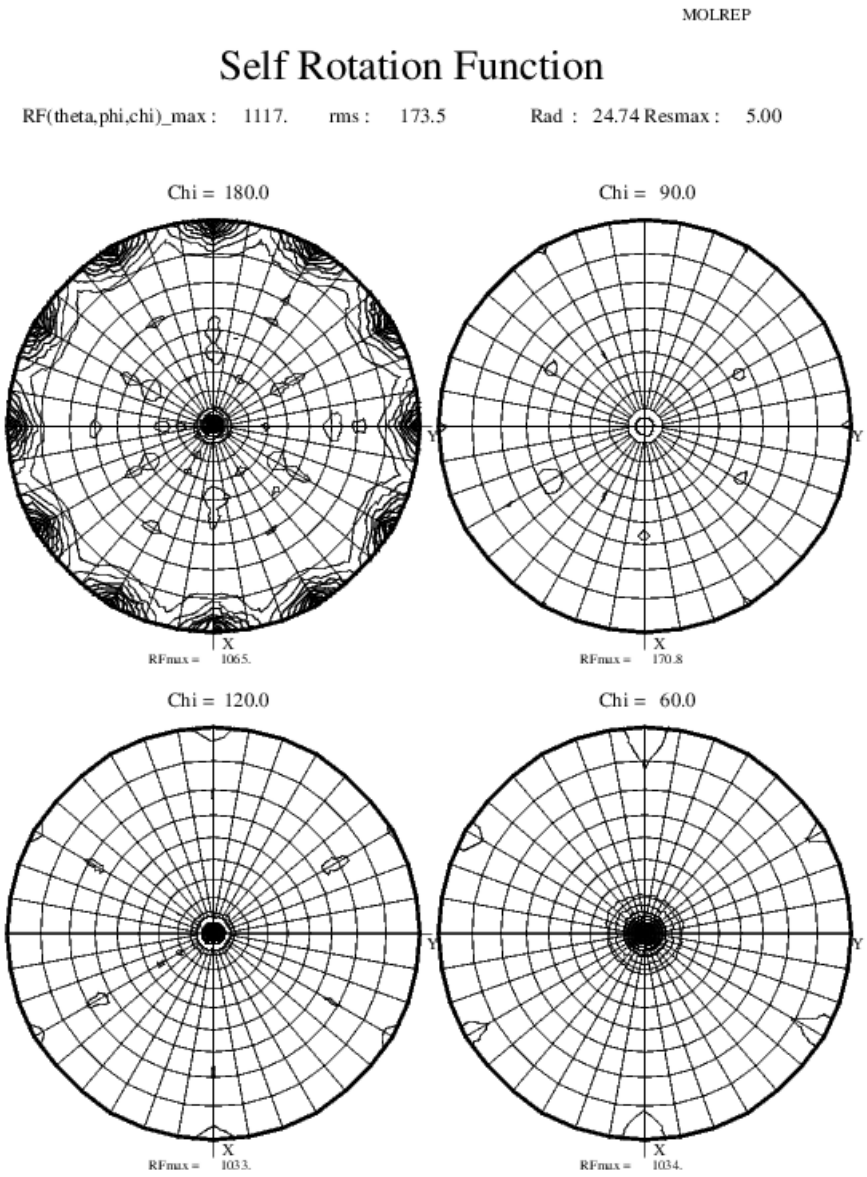



**Figure 5.10 Self Rotation Function of Data Set SeMet-Small**

The data used in the self-rotation function was processed in space group *P*1 and between 13-5 Å resolution using *MOLREP*.

**Table 5.4: ExeC Data Collection Statistics of Crystal Nat-Large**

| Nat-Large | | | |
|---|---|---|---|
| Space Group | *P*3 | $P6_5$ | $P6_522$ |
| Wavelength (Å) | 0.97949 | 0.97949 | 0.97949 |
| *a* | 48.0233 | 48.0233 | 48.0233 |
| *b* | 48.0233 | 48.0233 | 48.0233 |
| *c* | 301.3778 | 301.3778 | 301.3778 |
| γ | 120 | 120 | 120 |
| Resolution (Å) | 10.73-2.40(2.46-2.40) | 10.29-2.30(2.36-2.30) | 9.62-2.15(2.21-2.15) |
| All Reflections | 160164 (8828) | 171605(6966) | 185845(5231) |
| Unique Reflections | 30395(2137) | 17330(1237) | 12240 (876) |
| Multiplicity | 5.3 | 9.9 | 15.2 |
| Completeness (%) | 99.9 (99.3) | 99.9 (99.5) | 99.8 (99.4) |
| Mosaicity | 0.223 | 0.223 | 0.223 |
| I/Sigma(I) | 16.8(6.37) | 22.25(6.15) | 25.74 (4.58) |
| $R_{meas}$* | 7.4 (33.0) | 7.4(44.8) | 7.5 (51.4) |

These statistics are a result of processing diffraction images with *AutoProcess* at CMCF-ID with Friedels law set to True. *Rmeas (redundancy independent Rfactor (Diederichs & Karplus, 1997, *Nature Struct. Biol*. **4**, 269-275.)). Rmeas = Sum(h) [ Sqrt(n/(n-1)) Sum(j) [I(hj) - <I(hj)>] / Sum(hj).

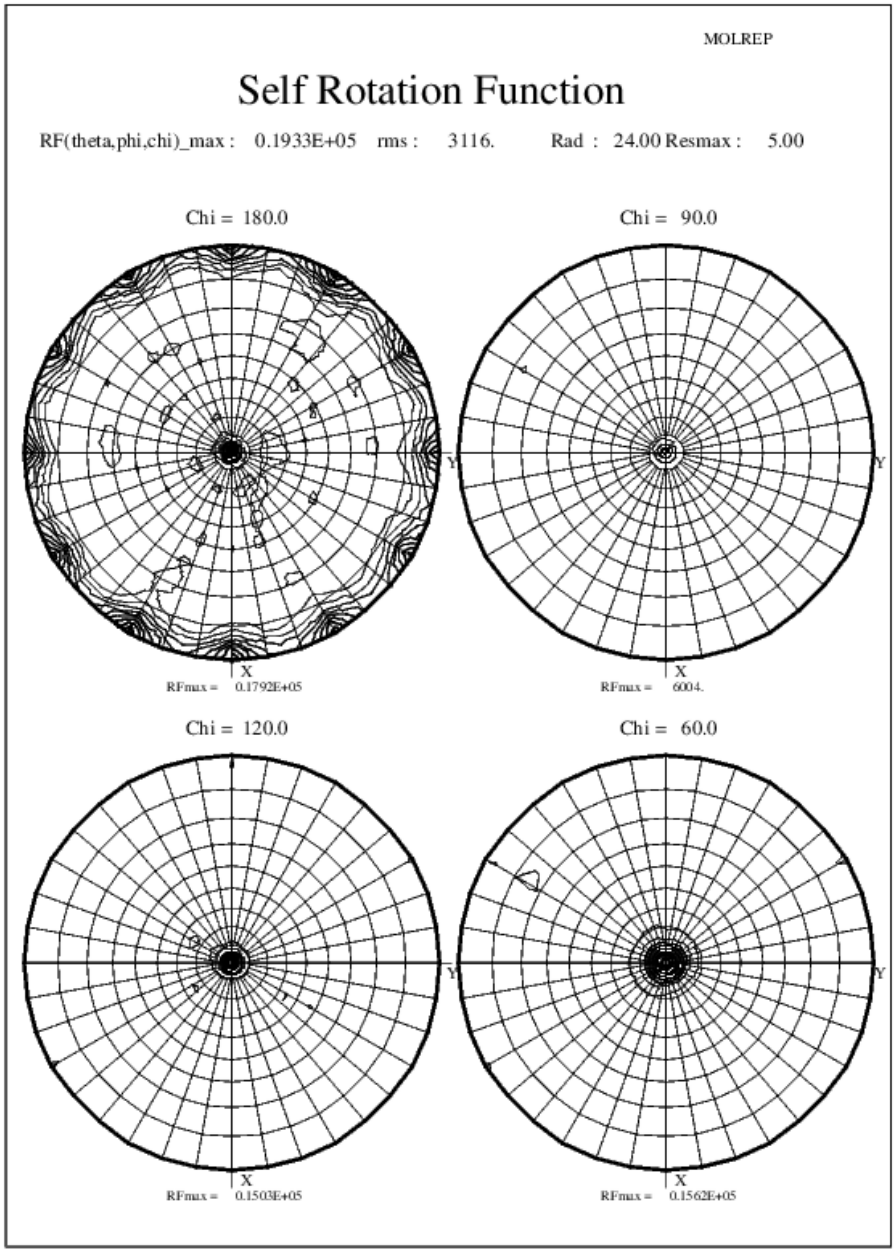



**Figure 5.11 Self Rotation Function of Data Set Nat-Large**

The data used in the self-rotation function was processed in space group *P*1 and between 13-5 Å resolution using *MOLREP*.

**Table 5.5:** List of (*0,0,l*) reflections for three Data Sets **(**Nat-Small, SeMet-Small and Nat-Large)

| Nat-Small *P*3 | | | | SeMet-Small *P*3 | | | | Nat-Large *P*3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Index *l* | I | sigI | I/sigI | Index *l* | I | sigI | I/sigI | Index *l* | I | sigI | I/sigI |
| 4 | -1 | 9 | -0.07 | 4 | 11 | 5 | 2.08 | | | | |
| 5 | 12 | 10 | 1.14 | 5 | 22 | 6 | 4.02 | | | | |
| 6 | 34367 | 487 | 70.55 | 6 | 4900 | 71 | 68.99 | | | | |
| 7 | -21 | 14 | -1.53 | 7 | 23 | 8 | 2.84 | 7 | 110 | 10 | 11.38 |
| 8 | 6 | 15 | 0.37 | 8 | 30 | 9 | 3.21 | 8 | 12 | 10 | 1.14 |
| 9 | 20418 | 291 | 70.17 | 9 | 1570 | 27 | 58.97 | 9 | 2 | 10 | 0.24 |
| 10 | 18 | 19 | 0.94 | 10 | 28 | 11 | 2.48 | 10 | 18 | 12 | 1.5 |
| 11 | 23 | 21 | 1.11 | 11 | -6 | 10 | -0.63 | 11 | -25748 | 395 | -65.17 |
| 12 | 126393 | 1792 | 70.54 | 12 | 35979 | 515 | 69.93 | 12 | 287980 | 4076 | 70.65 |
| 13 | -14 | 24 | -0.57 | 13 | 4 | 12 | 0.37 | 13 | -43215 | 615 | -70.29 |
| 14 | 6 | 25 | 0.22 | 14 | 0 | 13 | -0.01 | 14 | -8 | 17 | -0.44 |
| 15 | 142556 | 2021 | 70.55 | 15 | 24660 | 354 | 69.66 | 15 | -13 | 19 | -0.66 |
| 16 | -30 | 29 | -1.03 | 16 | -13 | 18 | -0.73 | 16 | 21 | 20 | 1.05 |
| 17 | -61 | 30 | -2.03 | 17 | 12 | 18 | 0.64 | 17 | 74 | 21 | 3.51 |
| 18 | 87914 | 1249 | 70.38 | 18 | 32327 | 461 | 70.07 | 18 | 45283 | 642 | 70.51 |
| 19 | 44 | 33 | 1.33 | 19 | 8 | 16 | 0.52 | 19 | 107 | 25 | 4.26 |
| 20 | -48 | 34 | -1.4 | 20 | -2 | 16 | -0.11 | 20 | 5605 | 86 | 65.46 |
| 21 | 94063 | 1341 | 70.14 | 21 | 4352 | 68 | 64.19 | 21 | 973 | 37 | 26.13 |
| 22 | -87 | 35 | -2.51 | 22 | -23 | 17 | -1.39 | 22 | 38 | 30 | 1.26 |
| 23 | -68 | 36 | -1.88 | 23 | 10 | 20 | 0.53 | 23 | 1555 | 39 | 39.47 |
| 24 | 65857 | 937 | 70.32 | 24 | 7382 | 111 | 66.55 | 24 | 353479 | 5002 | 70.66 |
| 25 | -58 | 40 | -1.47 | 25 | 19 | 22 | 0.85 | 25 | 932 | 39 | 23.64 |
| 26 | 104 | 41 | 2.53 | 26 | 32 | 22 | 1.47 | 26 | 5 | 35 | 0.15 |
| 27 | 37 | 42 | 0.9 | 27 | 89 | 21 | 4.24 | 27 | -35 | 35 | -1 |
| 28 | -22 | 43 | -0.5 | 28 | 18 | 20 | 0.91 | 28 | 90 | 36 | 2.48 |
| 29 | 32 | 45 | 0.72 | 29 | 20 | 23 | 0.87 | 29 | 269 | 40 | 6.7 |
| 30 | 36024 | 516 | 69.81 | 30 | 7193 | 110 | 65.34 | 30 | 438549 | 6205 | 70.68 |

Systematic absences as output by *POINTLESS.* Data sets Nat-Small and SeMet-Small both display the smaller unit cell. The systematic absences are observed at indices $l=3n$. This would suggest the presence of a screw axis and places the apparent symmetry as one of either $P6_222$ or $P6_422$. It should be noticed that *l* indices 27 and 42 (not shown) are expected, however they are very week for both, Nat-Small and SeMet-Small. For data set Nat-Large reflections are observed at indices $l=6n$ which suggest apparent symmetry of either $P6_122$ or $P6_522$ the match is not however perfect for example *l* indices 20 is unexpectedly present.

### 5.6.1 Anisotropic Correction

Anisotropic correction was attempted on highly anisotropic data (Figure 5.12).

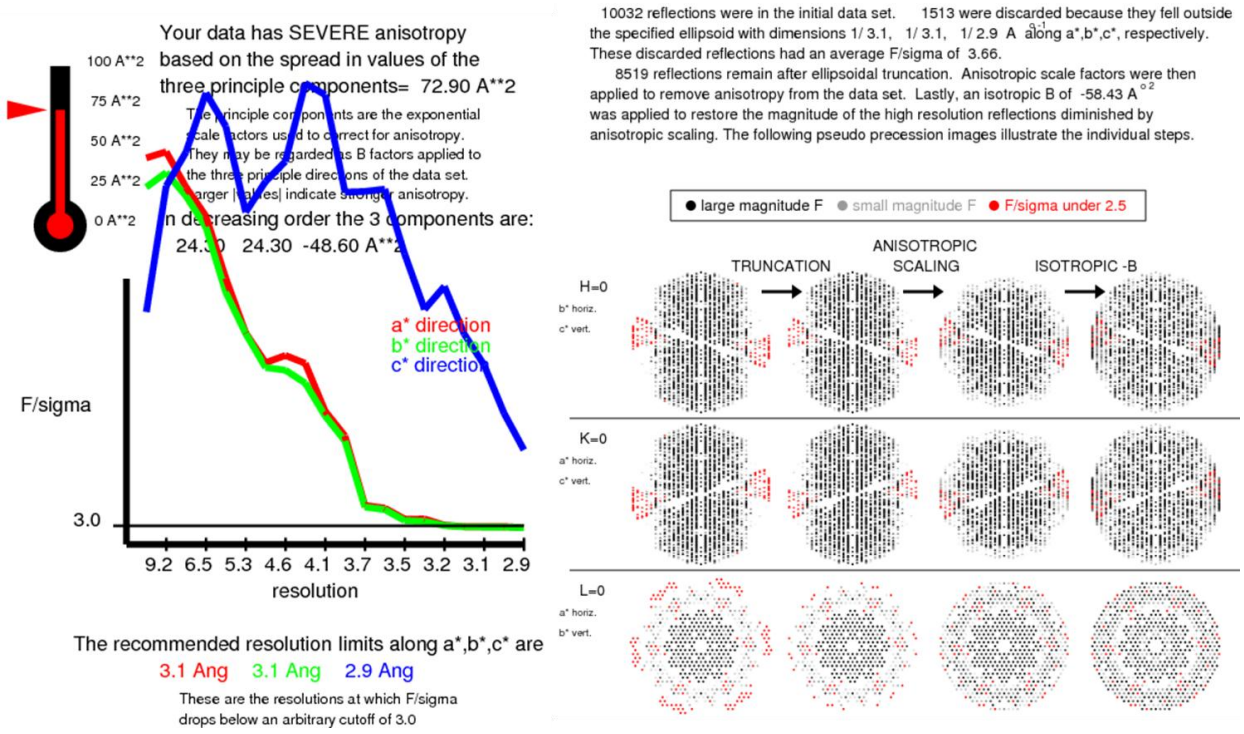



**Figure 5.12 UCLA Anisotropy Server Output**

This is an example of the output from the UCLA MBI — Diffraction Anisotropy Server. The data submitted was from the crystal Nat-Small in space group *P3*.

---

## 5.7 Twinning Analysis

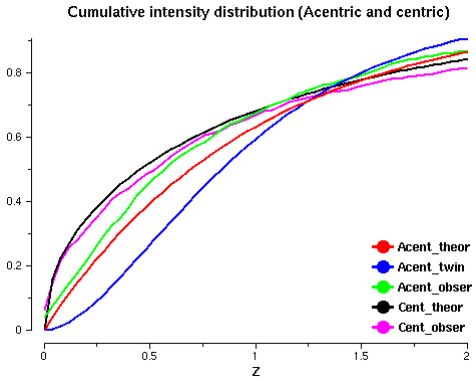
When intensity statistics were run on initial data sets the results did not look as they should for high quality untwinned data. All data submitted to twinning tests was processed using *AutoProcess* (CMCF) on reflection intensities.

### 5.7.1 Cumulative Intensity Distribution of *Z (N)Z*

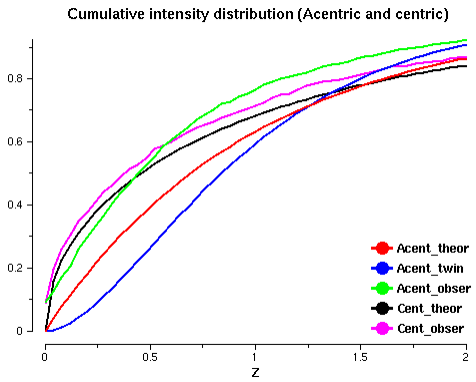
In Figure 4.13 A, the graph for observed acentric (green) and centric (pink) reflections are similar in shape to what is theoretically expected for untwined acentric (red) and centric (black) data. For acentric N(Z) values (Y-axis) between 0.25 Z and 1.5 Z are elevated. In Figure 4.13 B,

the plotted data follows the expected shape however for acentric data the N(Z) values are much higher than expected which is believed to be a result of anisotropy. In Figure 4.13 C, The observed acentric (green) and centric (pink) data similar in shape to that of theoretically twinned data (blue) indicating the possibility of twinning. Therefore, because the observed data does not look like twinned or untwinned one can conclude that this test does not provide any definitive answer.

A




B




C

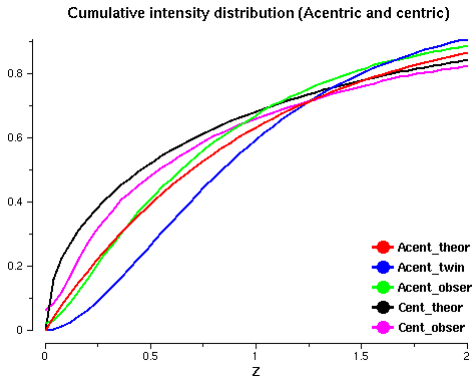



**Figure 5.13 Cumulative Intensity Distribution of Z (N)Z**

Intensity data was scaled and merged in space group $P6_422$ for small unit cell data (SeMet-Small (A) and Nat-Small (B)) and $P6_522$ for large unit cell data (Nat-Large (C)).

### 5.7.2 Perfect Twin Test

When the ‘truncate style’ twin test is conducted the small cell data has values much higher than what is expected for either twinned or untwinned data. Averaged values of three or higher are in these tests remain uninterpreted.

### 5.7.3 Partial Twinning Test

When the *H* test is carried out on ExeC data sets (data not shown) a twinning fraction of between 0.38 and 0.45 is consistently reported for all twin laws possible for triagonal space groups. A list of all possible twinning operators can be found in Appendix-4.

### 5.7.4 *L* Test

ExeC data looks normal in the large cell but in the small cell the acentric data displays the characteristics of centric data when the *L* test is applied (Figure 5.14). Centric data will have a concave slope with values below that expected for untwined data. This observation is true for both data sets SeMet-Small and Nat-Small. The observed data (red) for both of the small unit cell data sets fall below the line that is expected for untwined data. These results are more consistent with centric data and could be a result of pseudocentering. The results do not correspond to theoretical untwined data (blue) or theoretical perfectly twinned data (green). In the case of the large unit cell (Nat-Large) observed data does appear to be at least partially twinned.

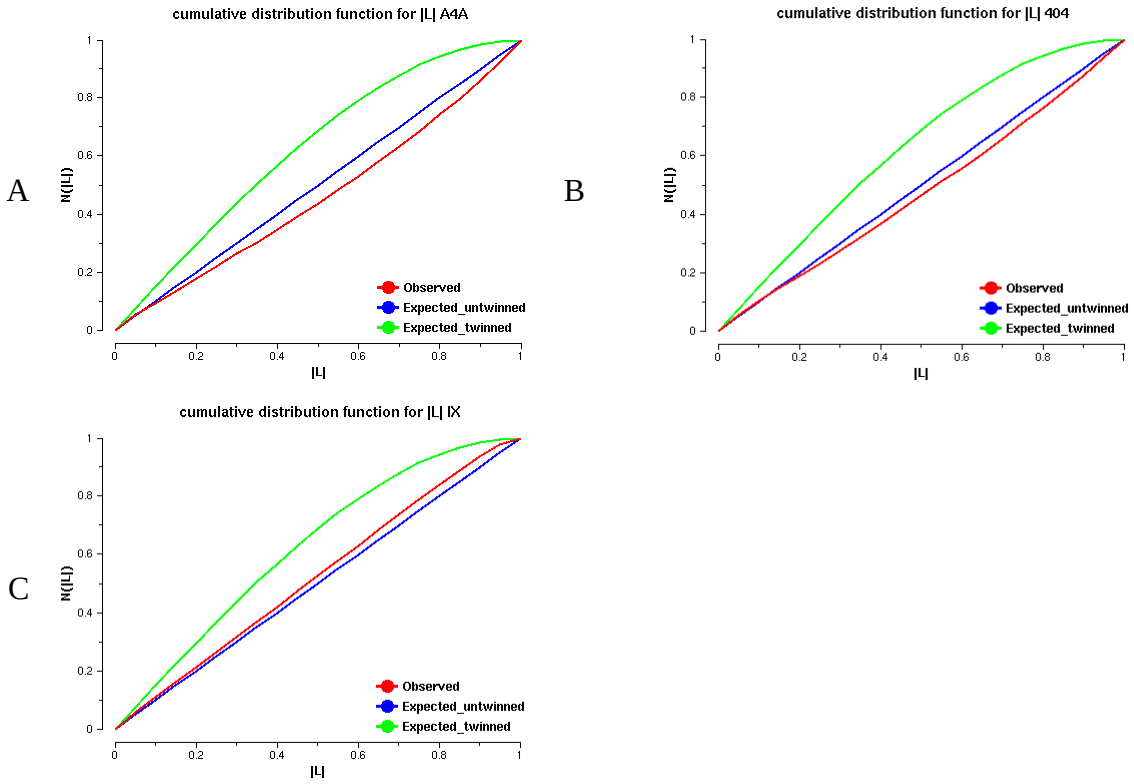



**Figure 5.14 *L* Test For Twinning**

Intensity data was scaled and merged in space group $P6_422$ for small unit cell data (SeMet-Small (A) and Nat-Small (B)) and $P6_522$ for large unit cell data (Nat-Large) (C).

---

## 5.8 Molecular Replacement on Data from Hexagonal ExeC Crystals

*PHASER (CCP4)* (Bailey 1994) produced a solution using dataset Nat-Large (Appendix-2). The crystal displayed the space group $P6_522$ / $P6_122$. The search model was the PDZ domain of EpsC edited by the program *CHAINSAW (CCP4)* (Bailey 1994). Unfortunately this solution could not be refined and the R-factors did not go below 48 % $R_{work}$ / 50 % $R_{free}$. Deleting loops and parts of the model that were not well shown by electron density map did not seem to improve the map nor did it give any clues as to how to build in the missing parts of the protein. Submitting the model and reflection file to the *PHENIX* (Adams *et al.* 2010) simulated annealing procedure did not improve the R-factors.

Subsequent data sets were processed with *HKL2000* in all possible point groups *P*1, *P*2, *C*2, *C*222, *P*3, *P*312, *P*321, *P*6 and *P*622. The data was subject to *TRUCNCATE* for ellipsoidal truncation and twinning analysis. An example of the *L* test performed by *TRUNCATE* is shown in (Figure 5.14). Then the data were submitted using all possible space groups (ex. *P*3 will search *P*3, $P3_1$ and $P3_2$) to the *BALBES* (Long *et al.* 2008) web server for molecular replacement. *BALBES* first analyzes the input reflection data by running *SFCHECK* (*CCP4*) (Bailey 1994). *BALBES* then extracts the unit-cell parameters, space group, data completeness, optimal resolution, the pseudo-translation vector if it exists, twin operators and estimates of the twin fraction. After *SFCHECK* (*CCP4*) (Bailey 1994) the input sequence is used to find homologues in the PDB in addition to 2I6V and 2I4S. There are three other models with lower sequence identity (1XED 26%, 2FB6 20.8%, 1S72 H 23.4%) that are used in the searches. Once the models have been chosen, molecular replacement commences.

The small cell data (SeMet-Small peak: appendix 2 & 3) produced a possible solution in space group $P3_212$ that has been refined to 47 % $R_{work}$ / 47 % $R_{free}$ with 50.7 % solvent content. This data also displays a SeMet anomalous signal which may be able to provide additional phase information. The most convincing evidence that the two solutions discussed above have some validity comes from inspection of the crystal packing. Both the big cell, in space group $P6_522$ and the small cell in, space group $P3_212$ have virtually the same crystal packing.

## 5.9 Attempts at MAD Phasing using Selenium in SeMet-ExeC Crystals

SeMet derivatized crystals were subject to an X-ray fluorescence scan above and below the K edge energy for Se, both to confirm the presence of Se in the crystal and to set the incident X-ray beam to the optimum wavelength (Figure 5.15). The crystallisable fragment is expected to have five Se atoms per monomer which should theoretically have plenty of phasing power for this size of protein. An anomalous signal was detected in the data collected on Se-containing crystals. All SeMet crystals displayed the small unit cell and appeared to have a strong anomalous signal. Initially all substructure solutions were of poor quality with low occupancies and no NCS when processed in lower space groups. Analyzing these sites did not reveal correlations to the MR solutions and the electron density maps were complete noise and not interpretable.

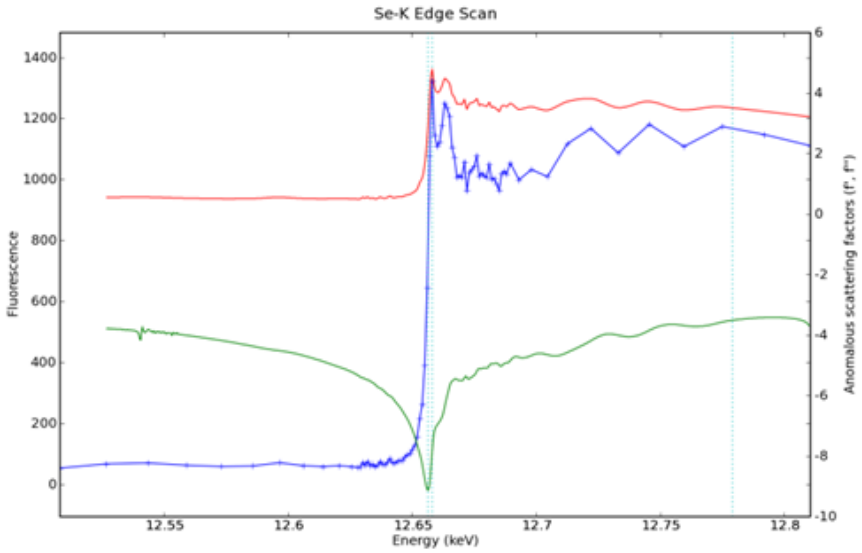



**Figure 5.15 X-ray Absorption Scan**

An example of a fluorescence scan collected at CLS 08ID-1 beamline. The blue trace is the observed fluorescence scan of Se while the green and red are the calculated f' and f" respectively, using the program CHOOCH (Evans and Pettifer 2001).

---

Eventually a substructure was found using the peak data set from crystal SeMet-Small in a newer version (2012) of *AutoSol* (*PHENIX*). This substructure contained four anomalous scattering sites in space group P6422 (Figure 5.16). Work on refining the structure based on these initial phases is ongoing.

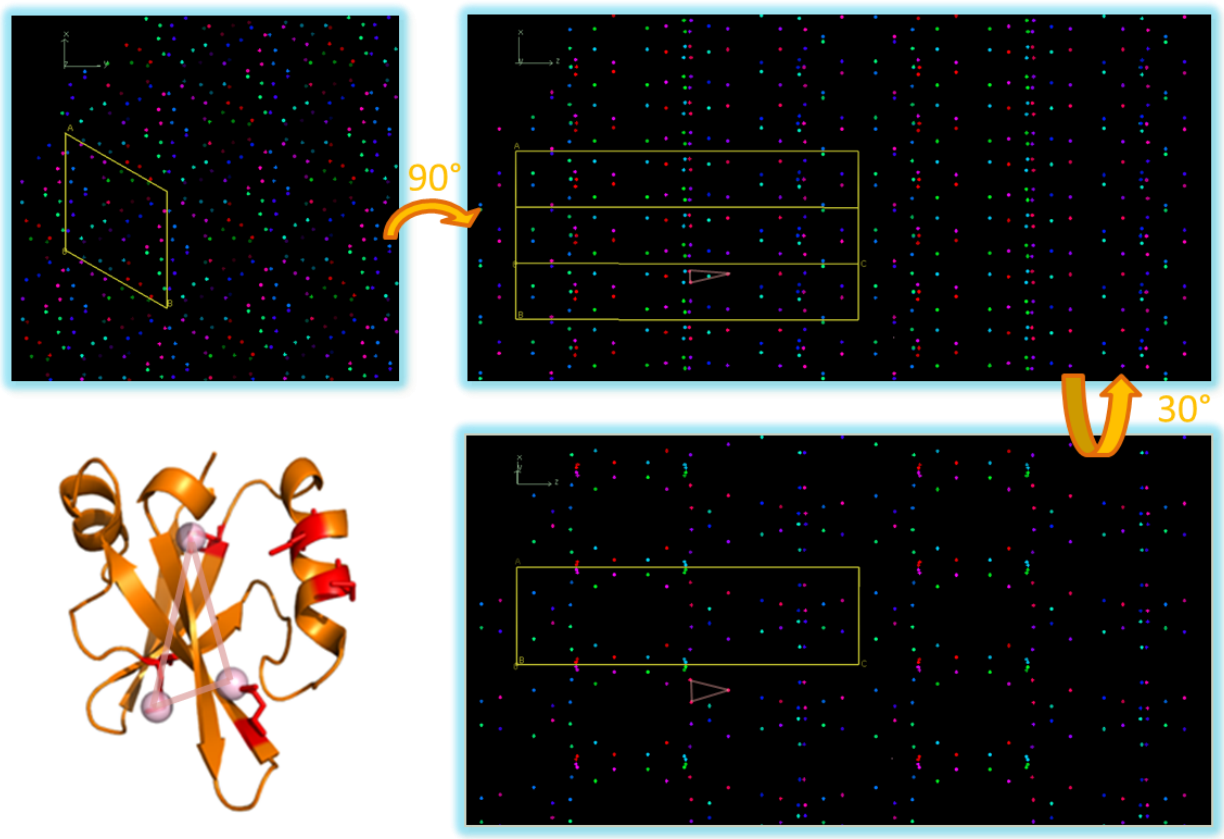



**Figure 5.16 Tentative Selenium Substructure**

An image looking at the x-z plane created using *COOT* shows the highest occupancy sites (red dots) found using *SHELX* in space group $P6_422$ (all other colors are symmetry related copies). The pink triangle connects three Se sites which correspond to a single protein monomer as shown in the ribbon diagram. The ribbon diagram shows the S atoms of the methionine (red) residues in the core of the PDZ domain from the MR search model derived from PDB 2I4S can be accurately superimposed with the substructure shown above (*PYMOL*).

# 6.0 DISCUSSION

ExeC is a very interesting specimen considering its critical role in the T2SS of *Aeromonas hydrophila*. The original construct, ExeC-ori, initially engineered for purposes other than crystallography was shown to be recalcitrant to crystallization. It is vulnerable to fragmentation and it was not designed for protein crystallography. The main problem is a 43 amino acid long tag on the amino terminus of the construct. The tag has unnecessary features such as the thrombin cleavage site, S•Tag and enterokinase cleavage site. These extra residues do not convey any biologically relevant secondary structure. Because it was immediately available, it was initially used for purification and crystallization studies. After it was determined that ExeC-ori was unstable two new constructs were designed specifically for this research. Overexpression of both ExeC HR-PDZ-C-tag and HR-C-tag was abundant and the over expressed protein could be readily captured through the use of a Ni-IMAC column.

In spite of significant efforts, ExeC-HR-PDZ-C-tag displayed similar properties to that of ExeC-ori in that it did not remain stable in the concentrated salt-free environment used for crystallization screening. ExeC-HR-PDZ also degraded into two subunits, although more slowly than ExeC-ori. HR-PDZ-C-tag did eventually yield a crystallisable ExeC fragment.

Growing and optimizing the growth of ExeC crystals was difficult because the generation of the crystallisable fragment was a sub-process that occurred in the crystallization drop and was not well understood. This process tended to take a very long time and the rate at which the crystallisable PDZ fragment generated was highly variable. It often took several months for crystals to form. By this time the protein solution used in the experiment was entirely used in additional crystallization trials or any remainder had undergone substantial changes in its fragmentation profile. This variability made optimization a very difficult task. Over time knowledge of ExeC crystal growth habits led to the genesis of well diffracting crystals.

SDS-PAGE analysis, with the aim of identifying the crystallizable fragment, using isolated crystals was not conclusive. It was not possible to wash them free of growth solution containing non-crystallizable ExeC fragments. Multiple bands would always appear. At this point the contents of the crystal could be; HR, PDZ, or HR-PDZ. Conclusive results on this issue were

finally obtained when a degraded ExeC protein sample was subsequently purified. Separating all species from one another was accomplished first by running the degraded sample over the Ni column. The HR domain would not bind and would be washed out of the Ni-IMAC column. The fragments left in the Ni-IMAC column were both the full length construct and the PDZ domain fragment with the tag. Once eluted, the binding fragments were loaded on the gel filtration column and separated. The reason that the separation was done in two steps was because the HR and PDZ fragments are too close in size to be resolved by gel filtration alone. All three ExeC samples were then submitted to crystallization trials and the PDZ fragment alone formed the familiar hexagonal Bi-pyramidal crystals thus confirming that the PDZ fragment was present in the crystals.

Getting a crystal from the crystallization drop to the goniometer is perhaps the most intense experience of the entire process. It requires manual manipulation of the precious specimen outside of its preferred environment (the vapour diffusion chamber). It is necessary to cool the crystal in liquid nitrogen in order to reduce radiation damage during data collection. This required the alteration of the solution within and around the crystal. The growth condition for ExeC crystals was near the saturation point of the ammonium sulphate precipitant. This was challenging because shortly after opening the vapour diffusion chamber the evaporation of water from the crystallization drop would cause the ammonium sulphate to crystallize in the drop. This often occurred while cryoprotection treatment was underway. ExeC crystals are believed to be very sensitive to changes in the drop condition. Side effects of cryoprotection were streaking of diffraction spots and unit cell doubling. One benefit of working with oil in conjunction with vapour diffusion drops is that the oil coats the drop and stops evaporation.

Few good quality diffraction data sets were collected. Effects of anisotropy, streaking spots and severe radiation damage corrupted the data quality of from most crystals. Even when reasonably good quality diffraction was observed there were often problems merging symmetry related reflections especially at resolution above 3 Å where the $R_{meas}$ values dramatically increased.

It is known that pseudo symmetry as well as anisotropic data can skew results from twinning tests. This could be why the *L* test and perfect twin test result in trends not consistent with

theoretical twinned or untwined values. *SFCHECK*, *DETWIN*, *TRUNCATE* (*CCP4*) and *XTRIAGE* (*PHENIX*) were used to implement the twinning analysis. All programs agree that some data sets are likely twinned. Twinning could account for the fact that data can be scaled and merged in essentially all hexagonal space groups with only small differences in $R_{meas}$. Evidence of twinning includes the intensity statistics test and also the observation that diffraction from lower quality crystals had split spots toward the diffracting limit of the crystal. This can happen when pseudo-merohedral twinning is present. There also appears to be translational NCS along c axis of the crystal which may at least partially hide twinning from intensity distribution tests. All of these factors make it difficult to deduce what type of crystal pathology is present.

In spite of these hurdles molecular replacement was attempted. Having a search model of greater than 40% identity was encouraging however no solution could be found. Tentative solutions still provide a reference point as to possible molecular packing configurations. These tentative solutions agree very well with the initial substructure solutions found using SAD phasing of SeMet crystals. This molecular packing appears to be amenable to twinning or lattice-translocation disorders.

Protein crystallography is a very diverse field that requires a unique combination of skills ranging from cell biology through genomics to high energy physics. Solving crystal structures requires state of the art technologies. Over the last decade, molecular techniques have allowed for proteins to be identified and expressed in high quantities in a number of expression vectors. Not only can the protein of interest be overexpressed it can be purified using highly specific tags for initial capture steps of purification. Although crystallization of protein molecules remains the bottleneck of a structural study, it too has succumbed to advancement, whether it be high-throughput automated screening using sub-microliter volumes of protein or the use of oils and temperature blocks to make systematic studies of the effects of temperature.

On the hardware end of data collection, the ultimate tool for a crystallographer is the synchrotron. From undulator to X-ray detector a crystallography beamline is truly a marvel of technology. The synchrotron allows the crystallographer to push the limits of phasing and resolution. The monochromatic beam allows for SAD/MAD phasing sometimes from native

crystals using anomalous differences from sulfur. Not only does a synchrotron aid in phasing but the high flux can push the resolution limit which in turn elucidates the chemical details inherent in proteins and critical to their functions. This after all is the end goal.

## 7.0 References

Abendroth, J., M. Bagdasarian, *et al*. (2004). "The structure of the cytoplasmic domain of EpsL, an inner membrane component of the type II secretion system of Vibrio cholerae: An unusual member of the actin-like ATPase superfamily." Journal of Molecular Biology 344(3): 619-633.

Abendroth, J., A. C. Kreger, *et al*. (2009). "The dimer formed by the periplasmic domain of EpsL from the Type 2 Secretion System of Vibrio parahaemolyticus." Journal of Structural Biology 168(2): 313-322.

Abendroth, J., D. D. Mitchell, *et al*. (2009). "The three-dimensional structure of the cytoplasmic domains of EpsF from the type 2 secretion system of Vibrio cholerae." Journal of Structural Biology 166(3): 303-315.

Abendroth, J., P. Murphy, *et al*. (2005). "The X-ray structure of the type II secretion system complex formed by the N-terminal domain of EpsE and the cytoplasmic domain of EpsL of Vibrio cholerae." Journal of Molecular Biology 348(4): 845-855.

Abendroth, J., A. E. Rice, *et al*. (2004). "The crystal structure of the periplasmic domain of the type II secretion system protein EpsM from Vibrio cholerae: The simplest version of the ferredoxin fold." Journal of Molecular Biology 338(3): 585-596.

Adams, P. D., P. V. Afonine, *et al*. (2010). "PHENIX: a comprehensive Python-based system for macromolecular structure solution." Acta Crystallographica Section D-Biological Crystallography 66: 213-221.

Akrim, M., M. Bally, *et al*. (1993). "XCP-mediated protein secretion in pseudomonas-aeruginosa - identification of 2 additional genes and evidence for regulation of xcp gene-expression." Molecular Microbiology 10(2): 431-443.

Ast, V. M., I. C. Schoenhofen, *et al*. (2002). "Expression of the ExeAB complex of Aeromonas hydrophila is required for the localization and assembly of the ExeD secretion port multimer." Molecular Microbiology 44(1): 217-231.

Bailey, S. (1994). "The *CCP4* suite: programs for protein crystallography." Acta Crystallographica Section D-Biological Crystallography 50: 760-763.

Bleves, S., M. Gerard-Vincent, *et al*. (1999). "Structure-function analysis of XcpP, a component involved in general secretory pathway-dependent protein secretion in Pseudomonas aeruginosa." Journal of Bacteriology 181(13): 4012-4019.

Brok, R., P. Van Gelder, *et al*. (1999). "The C-terminal domain of the Pseudomonas secretin XcpQ forms oligomeric rings with pore activity." Journal of Molecular Biology 294(5): 1169-1179.

Brunger, A. T. (1991). "Simulated Annealing in Crystallography." Annual Review of Physical Chemistry 42(1): 197-223.

Chenna, R., H. Sugawara, *et al*. (2003). "Multiple sequence alignment with the Clustal series of programs." Nucleic Acids Research 31(13): 3497-3500.

Collins, R. F., S. A. Frye, *et al*. (2004). "Structure of the Neisseria meningitidis outer membrane PilQ secretin complex at 12 angstrom resolution." Journal of Biological Chemistry 279(38): 39750-39756.

Dauter, Z. (2003). "Twinned crystals and anomalous phasing." Acta Crystallographica Section D 59(11): 2004-2016.

Doublie, S. (1997). "Preparation of selenomethionyl proteins for phase determination." Macromolecular Crystallography, Pt A 276: 523-530.

Douet, V., L. Loiseau, *et al*. (2004). "Systematic analysis, by the yeast two-hybrid, of protein interaction between components of the type II secretory machinery of Erwinia chrysanthemi." Research in Microbiology 155(2): 71-75.

Douzi, B., E. Durand, *et al*. (2009). "The XcpV/GspI Pseudopilin Has a Central Role in the Assembly of a Quaternary Complex within the T2SS Pseudopilus." Journal of Biological Chemistry 284(50): 34580-34589.

Evans, G. and R. F. Pettifer (2001). "CHOOCH: a program for deriving anomalous-scattering factors from X-ray fluorescence spectra." Journal of Applied Crystallography 34(1): 82-86.

Filloux, A. (2004). "The underlying mechanisms of type II protein secretion." Biochimica Et Biophysica Acta-Molecular Cell Research 1694(1-3): 163-179.

Forest, K. T. (2008). "The type II secretion arrowhead: the structure of GspI-GspJ-GspK." Nature Structural & Molecular Biology 15(5): 428-430.

Gu, S., G. Kelly, *et al*. (2012). "Solution Structure of Homology Region (HR) Domain of Type II Secretion System." Journal of Biological Chemistry 287(12): 9072-9080.

Hazes, B. and L. Frost (2008). "Towards a systems biology approach to study type II/IV secretion systems." Biochimica Et Biophysica Acta-Biomembranes 1778(9): 1839-1850.

Howard, S. P., J. Critch, *et al*. (1993). "Isolation and analysis of 8 exe genes and their involvement in extracellular protein secretion and outer-membrane assembly in Aeromonas-hydrophila." Journal of Bacteriology 175(20): 6695-6703.

Jiang, B. and S. P. Howard (1992). "The Aeromonas-hydrophila *exe*E gene, required both for protein secretion and normal outer membrane biogenesis, is a member of a general secretion pathway." Molecular Microbiology 6(10): 1351-1361.

Johnson, T. L., J. Abendroth, *et al*. (2006). "Type II secretion: from structure to function." Fems Microbiology Letters 255(2): 175-186.

Kabsch, W. (1993). "Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants." Journal of Applied Crystallography 26: 795-800.

Koch, E. (1992). International Tables for Crystallography, Dordrecht: Kluwar Academic Publishers.

Korotkov, K. V., M. D. Gray, *et al*. (2009). "Calcium Is Essential for the Major Pseudopilin in the Type 2 Secretion System." Journal of Biological Chemistry 284(38): 25466-25470.

Korotkov, K. V. and W. G. J. Hol (2008). "Structure of the GspK-GspI-GspJ complex from the enterotoxigenic Escherichia coli type 2 secretion system." Nature Structural & Molecular Biology 15(5): 462-468.

Korotkov, K. V., T. L. Johnson, *et al*. (2011). "Structural and Functional Studies on the Interaction of GspC and GspD in the Type II Secretion System." PLoS Pathog 7(9): e1002228.

Korotkov, K. V., B. Krumm, *et al*. (2006). "Structural and functional studies of EpsC, a crucial component of the type 2 secretion system from Vibrio cholerae." Journal of Molecular Biology 363(2): 311-321.

Korotkov, K. V., E. Pardon, *et al*. (2009). "Crystal Structure of the N-Terminal Domain of the Secretin GspD from ETEC Determined with the Assistance of a Nanobody." Structure 17(2): 255-265.

Lam, A. Y., E. Pardon, *et al*. (2009). "Nanobody-aided structure determination of the EpsI:EpsJ pseudopilin heterodimer from Vibrio vulnificus." Journal of Structural Biology 166(1): 8-15.

Lee, H. M., K. C. Wang, *et al*. (2000). "Association of the cytoplasmic membrane protein XpsN with the outer membrane protein XpsD in the type II protein secretion apparatus of Xanthomonas campestris pv. campestris." Journal of Bacteriology 182(6): 1549-1557.

Long, F., A. A. Vagin, *et al*. (2008). "BALBES: a molecular-replacement pipeline." Acta Crystallographica Section D-Biological Crystallography 64: 125-132.

Otwinowski, Z. and W. Minor (1997). "Processing of X-ray diffraction data collected in oscillation mode." Macromolecular Crystallography, Pt A 276: 307-326.

Padilla, J. E. and T. O. Yeates (2003). "A statistic for local intensity differences: robustness to anisotropy and pseudo-centering and utility for detecting twinning." Acta Crystallographica Section D-Biological Crystallography 59: 1124-1130.

Pflugrath, J. (1999). "The finer things in X-ray diffraction data collection." Acta Crystallographica Section D 55(10): 1718-1725.

Possot, O. M., G. Vignon, *et al*. (2000). "Multiple interactions between pullulanase secreton components involved in stabilization and cytoplasmic membrane association of PulE." Journal of Bacteriology 182(8): 2142-2152.

Pugsley, A. P. (1993). "The complete general secretory pathway in gram-negative bacteria." Microbiological Reviews 57(1): 50-108.

Rees, D. (1982). "A general theory of X-ray intensity statistics for twins by merohedry." Acta Crystallographica Section A 38(2): 201-207.

Rees, D. C. (1980). "The influence of twinning by merohedry on intensity statistics." Acta Crystallographica Section A 36(JUL): 578-581.

Robert, V., F. Hayes, *et al*. (2002). "Identification of XcpZ domains required for assembly of the secreton of Pseudomonas aeruginosa." Journal of Bacteriology 184(6): 1779-1782.

Robien, M. A., B. E. Krumm, *et al*. (2003). "Crystal structure of the extracellular protein secretion NTPase EpsE of Vibrio cholerae." Journal of Molecular Biology 333(3): 657-674.

Sandkvist, M., J. M. Keith, *et al*. (2000). "Two Regions of EpsL Involved in Species-Specific Protein-Protein Interactions with EpsE and EpsM of the General Secretion Pathway in Vibrio cholerae." J. Bacteriol. 182(3): 742-748.

Sandkvist, M., L. O. Michel, *et al*. (1997). "General secretion pathway (eps) genes required for toxin secretion and outer membrane biogenesis in Vibrio cholerae." Journal of Bacteriology 179(22): 6994-7003.

Schoenhofen, I. C., C. Stratilo, *et al*. (1998). "An ExeAB complex in the type II secretion pathway of Aeromonas hydrophila: effect of ATP-binding cassette mutations on complex formation and function." Molecular Microbiology 29(5): 1237-1247.

Stanley, E. (1955). "The distribution of the mean intensity of a finite group of reflexions." Acta Crystallographica 8(6): 351-352.

Stanley, E. (1972). "Identification of twins from intensity statistics." Journal of Applied Crystallography 5(JUN1): 191-&.

Terwilliger, T. C., R. W. Grosse-Kunstleve, *et al*. (2008). "Iterative-build OMIT maps: map improvement by iterative model building and refinement without model bias." Acta Crystallographica Section D 64(5): 515-524.

Tsai, R. T., W. M. Leu, *et al*. (2002). "A reversibly dissociable ternary complex formed by XpsL, XpsM and XpsN of the Xanthomonas campestris pv. campestris type II secretion apparatus." Biochemical Journal 367: 865-871.

Waterhouse, A. M., J. B. Procter, *et al*. (2009). "Jalview Version 2—a multiple sequence alignment editor and analysis workbench." Bioinformatics 25(9): 1189-1191.

Yamagata, A. and J. A. Tainer (2007). "Hexameric structures of the archaeal secretion ATPase GspE and implications for a universal secretion mechanism." Embo Journal 26(3): 878-890.

Yanez, M. E., K. V. Korotkov, *et al*. (2008). "The crystal structure of a binary complex of two pseudopilins: EpsI and EpsJ from the type 2 secretion system of Vibrio vulnificus." Journal of Molecular Biology 375(2): 471-486.

Yanez, M. E., K. V. Korotkov, *et al*. (2008). "Structure of the minor pseudopilin EpsH from the type 2 secretion system of Vibrio cholerae." Journal of Molecular Biology 377(1): 91-103.

# 8.0 APPENDICES

## 8.1 Appendix-1 Crystallization Trials

| Protein | ID | commercial screen | # of plates | # of drops/well | # of conditions |
|---|---|---|---|---|---|
| long | 1-24 AmSo4 50% + 100 mM BT pH 6.5 Aug 19th | 50% AmSO4 (Nextal®) | 1 | 1 | 24 |
| long | A4 July 10 (2006) butterfly opti | | 1 | 1 | 24 |
| long | A3 | | 1 | 1 | 24 |
| long | A5 | | 1 | 1 | 24 |
| long | Dec 11 Nov 10th protein 20 mg/ml room/4/37 degree | | 1 | 3 | 72 |
| long | Dec 11 Malonate | | 1 | 1 | 24 |
| long | Jan 29th 08 pH Clear | | 4 | 2 | 192 |
| long | consortium Room temp June 7th 08 | | 1 | 1 | 24 |
| long | Jan 29th 08 AmSO4 X 4 | AmSO4 (Nextal®) | 4 | 2 | 192 |
| long | Feb 05 2009 opti | | 1 | 1 | 24 |
| long | June 7th opti long | | 1 | 1 | 24 |
| long | Nov 16 opti | | 1 | 1 | 24 |
| long | Aug 18th 1 M AmSO4 ADDITIVE X 4 | | 4 | 1 | 96 |
| long | 0.1M BT pH x 1 M AmSO4 1% PEG 3350 | | 1 | 1 | 24 |
| long | Mar 5th pH adjusted AmSO4 | | 1 | 1 | 24 |
| long | April 1st JCSG X4 | JCSG (Nextal®) | 4 | 2 | 192 |
| long | Feb 10th Opti | | 1 | 1 | 24 |
| long | May 28 April 17th protein Opti | | 1 | 1 | 24 |
| long | Opti X 4 pH 5/5.5/6/6.5 Oct 12 mono-Q two drop/well | | 4 | 2 | 192 |
| long | Nov 16 opti | | 1 | 1 | 24 |
| long | Nov 16 opti Ni-Tri | | 1 | 1 | 24 |
| long | Sept 18 o7 JCSG # 86 opti room | | 1 | 1 | 24 |
| long | Sept 18 o7 JCSG # 86 opti 4 deg | | 1 | 1 | 24 |
| long | PEG 10 K 2-5% April 17 | | 1 | 1 | 24 |
| long | PEG 10 K 2-15% x NaCl pH 6.5 April 17 | | 1 | 1 | 24 |
| long | PEG 2-18% April 17 | | 1 | 1 | 24 |
| long | PEG 10 K 2-15% x NaCl pH 7.5 April 17 | | 1 | 1 | 24 |
| long | April 2nd salts | | 1 | 1 | 24 |
| long | Oct 22 opti | | 1 | 1 | 24 |
| long | comPAS 2 and 3 | comPAS (Nextal®) | 2 | 2 | 96 |
| long | Jan 29th PEGS X 4 | PEG (Nextal®) | 4 | 2 | 192 |
| long | Feb 2nd JCSG + 150 mM NaCl X 4 | JCSG (Nextal®) | 4 | 1 | 96 |
| ori | MgCl2 PEG 400 grid Sept 22 06 | | 1 | 1 | 24 |
| short | Jan 29th pH Clear | | 4 | 2 | 192 |
| short | Jan 29th 08 AmSO4 X 4 | AmSO4 (Nextal®) | 4 | 2 | 192 |
| short | April 1st JCSG X4 | JCSG (Nextal®) | 4 | 2 | 192 |
| short | Aug 19 MPD pH 7.5 X 4 | MPD (Nextal®) | 4 | 1 | 96 |
| short | Aug 20 short classic X4 | Classic (Nextal®) | 4 | 1 | 96 |
| short | comPAS 2 and 3 | comPAS (Nextal®) | 2 | 2 | 96 |
| short | Jan 29th PEGS X 4 | | 4 | 2 | 192 |
| | no markings' | | 1 | 1 | 24 |
| ? | April 10th PACT X 4 | PACT (Nextal®) | 4 | 1 | 96 |
| ? | Mar 19 PEG II X 4 | PEG II (Nextal®) | 4 | 1 | 96 |
| ? | Jan 27th Classic 3 and 4 | Classic (Nextal®) | 2 | 1 | 48 |
| beta-D | May beta-D 16 2008 | | 1 | 1 | 24 |
| beta-D | feb 11 50% AmSO4 25-48 | | 1 | 1 | 24 |
| beta-D | May 22 08 opti | | 1 | 1 | 24 |
| beta-D | Mar 16th | | 1 | 1 | 24 |
| beta-D | Feb 26 2008 1-24 classic | | 1 | 1 | 24 |
| beta-D | B-D 0.6M AmSO4 1.5% isopropanol ADDITIVE X 4 | ADDITIVE (Hampton®) | 4 | 1 | 96 |
| beta-D | May 16 2008 opti | | 1 | 1 | 24 |
| beta-D | pH Clear Jan 6 09 1 and 4 | pH CLEAR (Nextal®) | 2 | 1 | 48 |
| | | | | **Total:** | 3504 |

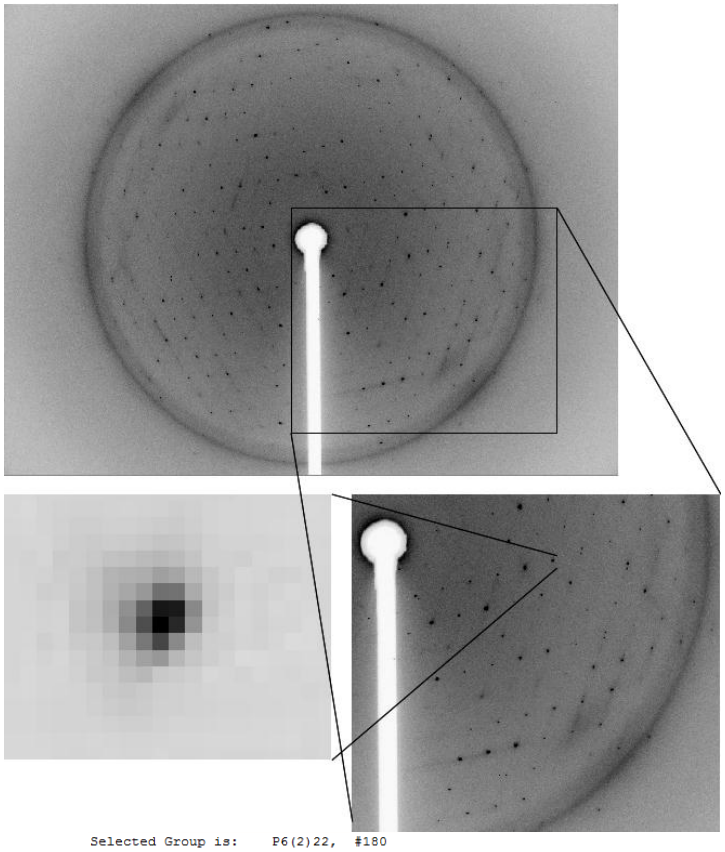
## 8.2 Appendix-2 Table of all Data Sets Collected

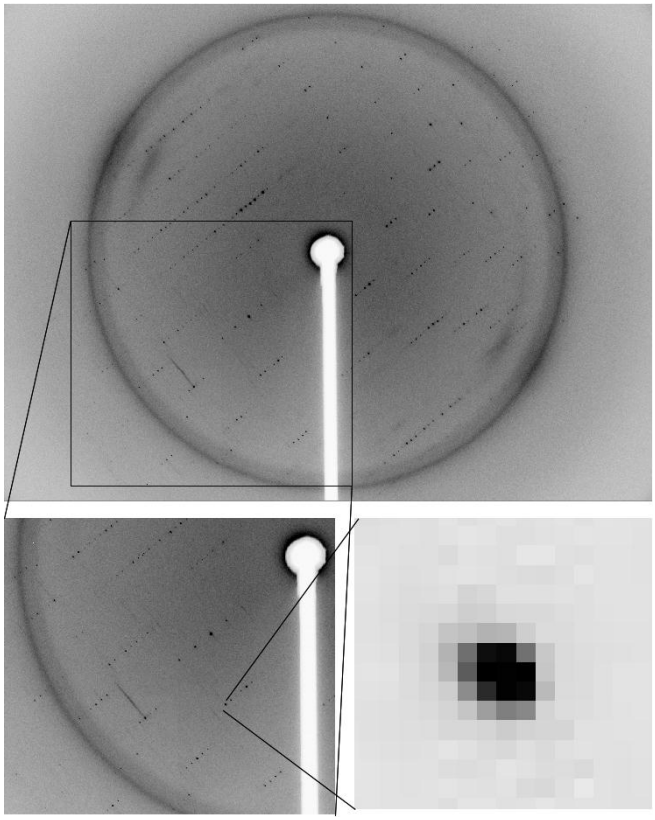
(SeMet-Small – A4A, Nat-Small – 404, Nat-Large – IX)

| Xtal ID | Date collected | Wavelength (Å) | Space Group | a | b | c | XDS crystal score | I/σ > 1 (Å) | I/σ > 3 (Å) | Total number of observations | Total [highest res shell] | Completeness [%] | R_merge [total] | R_meas (I+/-) [total] | Mosaicity [Osc. Range °] | Number of images | Growth condition | Heavy Atom | Cryo | Note |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| benmar7b5 | 07.03.2008 | 0.97934 | P6322 | 53.16 | 53.16 | 149.65 | ? | 3.05 | 3.70 | 68639 | 77.9 [3.44] | 24.7 | 22.6 | 0.41 | 180 [1.0] | Nov 27th plate well B5 | | Native | 30% Gly in' Nov27th' C5 well solution | |
| benmar7c5 | 07.03.2008 | 0.97934 | P6422 | 52.41 | 52.41 | 148.93 | ? | 3.20 | 3.60 | 29983 | 78.5 [3.48] | 15.9 | 17.8 | 0.31 | 180 [1.0] | Nov 27th' plate well C5 | | Native | 30% Gly in' Nov27th' C5 well solution | |
| benmar8d4 | 08.03.2008 | 0.97934 | | | | | | | | | | | | | 180 [1.0] | | | Native | | very poor data set |
| benmar13v1 | 13.03.2008 | 0.97934 | P6422 | 52.61 | 52.61 | 149.13 | ? | 2.65 | 2.95 | 103477 | 77.2 [2.70] | 26.0 | 16.4 | 0.29 | 720 [0.5] | Nov 27th' late ?? (b3) | | Native | likely Gly ?? | |
| benmar13v3 | 13.03.2008 | 0.97934 | P6422 | 53.14 | 53.14 | 149.28 | ? | 3.05 | 4.00 | 22975 | 74.4 [3.59] | 19.2 | 29.0 | 0.36 | 180 [0.5] | Nov 27th' late ?? (b3) | | Native | likely Gly ?? | |
| 404 | 30.07.2008 | 0.97934 | P6222 | 52.57 | 52.57 | 150.30 | 0.646 | <2.50 | 2.60 | 31333 | 81.4 [2.53] | 6.3 | 11.8 | 0.21 | 80 [1.0] | 0.75M AmSo 1%400 pH5.4 ' Jan27th' plate | | Native | ? | cell dimensios for P3 SG |
| btob | 30.07.2008 | 0.97934 | C121 | 52.50 | 52.50 | 150.59 | 0.272 | 2.53 | 3.20 | 27455 | 58.3 [2.82] | 9.8 | 24.5 | 0.46 | 100 [1.0] | Jan27th plate 0.75M AmSo pH5.4 | | Native | 40% PEG 400 | |
| O3 | 05.09.2008 | 0.97934 | P6522 | 48.70 | 48.70 | 302.80 | 0.271 | 2 | 2.65 | 210904 | 86.1 [2.34] | 22.4 | 18.1 | 0.15 | 180 [1.0] | 0.9M AmSo 100 mM BT 5.6 | | Native | 40% PEG 400 | |
| 10 | 07.09.2008 | 0.97934 | P6522 | 53.61 | 53.61 | 297.82 | 0.678 | 2.00 | 2.55 | 64680 | 76.7 [2.20] | 9.2 | 20.8 | 0.13 | 101 [0.5] | 0.8M AmSo 100 mM BT 5.4 | | Native | 18% Glycerol | |
| O6 | 10.09.2008 | 1.07225 | C121 | 48.69 | 48.78 | 302.30 | 0.233 | 2.85 | 3.90 | 84105 | 77.4 [3.58] | 27.5 | 37.1 | 0.27 | 160 [1.0] | 0.9M AmSo 100 mM BT 5.6 | | Au | 30% PEG 400 | cell dimensios for P3 SG |
| 15 | 21.09.2008 | 1.48023 | P622 | 48.56 | 48.56 | 302.18 | 0.447 | 2.45 | 3.25 | 54058 | 79.5 [2.66] | 15.9 | 27.2 | 0.26 | 283 [1.0] | 0.8M AmSo 100 mM BT 5.2 | | Gd | 40% PEG 400 | |
| PtP1 | 21.09/2008 | 1.07225 | C2 | 48.40 | 48.40 | 302.50 | | 2.75 | 4.00 | 55149 | 69.2 [3.03] | 44.8 | 47.1 | 0.25 | 360 [0.5] | | | Pt | | cell dimensios for P3 SG |
| | | 1.07225 | P6122 | 48.56 | 48.56 | 303.60 | 0.443 | 2.45 | 2.79 | 205071 | 91.9 [2.55] | 18.9 | 10.5 | 0.26 | 567 [0.5] | | | Pt | | |
| PtP2 | 22.09.2008 | 1.07225 | C222 | 48.50 | 48.50 | 302.70 | 0.432 | 2.4 | 3.10 | 140135 | 90.2 [2.84] | 19.7 | 23.6 | 0.19 | 180 [1.0] | | | Pt | | cell dimensios for P3 SG |
| 1a | 13.12.2008 | 1.48023 | P6122 | 48.24 | 48.24 | 303.40 | 0.427 | 2.15 | 2.69 | 200837 | 89.1 [2.43] | 17.7 | 16.7 | 0.36 | 360 [1.0] | Nov27th plate well A2 | | Native | 70% Malonate | |
| 4a | 13.12.2008 | 1.48023 | P6122 | 49.00 | 49.00 | 302.99 | 0.428 | 2.55 | 3.20 | 40296 | 89.5 [3.00] | 10.4 | 14.5 | 0.47 | 180 [1.0] | Nov27th plate well A2 ? Or June 6th NaCl B1 | | Gd | 70% Malonate | Statistics on 114 of 180 img 1hrGd soak with backwashing |
| 6a | 13.12.2008 | 1.48023 | P6122 | 48.38 | 48.38 | 303.85 | 0.506 | 1.95 | 2.75 | 98295 | 70.1 [2.11] | 11.8 | 27.4 | 0.39 | 180 [1.0] | Nov27th plate well A2 | | Gd | 70% Malonate | 15 min soak + backwashing |
| 6d | 13.12.2008 | 1.48023 | P6122 | 48.56 | 48.56 | 303.36 | 0.431 | 2.5 | 3.20 | 72663 | 84.8 [2.74] | 13.9 | 27.2 | 0.43 | 180 [1.0] | Nov27th plate well A2 | | Gd | 70% Malonate | 1.5 hr. soak |
| B3 | 14.12.2008 | 1.48023 | P6122 | 48.84 | 48.84 | 302.82 | 0.542 | 2.24 | 2.80 | 119244 | 87.6 [2.54] | 11.0 | 14.9 | 0.28 | 180 [1.0] | Nov27th plate well A2 | | Gd | 70% Malonate | 2 hr. Soak + backwash through 3 drops |
| 3 | 25.01.2009 | 1.48129 | P6222 | 48.83 | 48.83 | 151.41 | 0.534 | <2.79 | 2.97 | 39609 | 88.7 [2.79] | 8.8 | 7.7 | 0.20 | 180 [1.0] | | | Gd | | |
| 6c | 25.01.2009 | 1.48129 | C222 | 48.30 | 48.30 | 151.20 | 0.061 | 2.80 | 3.38 | 18584 | 80.5 [3.16] | 26.7 | 32.0 | 0.48 | 180 [0.5] | | | Gd | | cell dimensios for P3 SG |
| 8b | 25.01.2009 | | | | | | not | processed | | | | | | | 90 [?.?] | | | Gd | | streaky spots multiple sattices |
| 79 | 25.01.2009 | | | | | | | HKL2000 | only | | | | | | 120 [?.?] | | | Gd | | |
| B4 | 25.01.2009 | 1.48129 | P6222 | 48.59 | 48.59 | 151.21 | 0.466 | <2.68 | 3.02 | 16632 | 87.5 [2.84] | 8.7 | 11.7 | 0.27 | 106 [1.0] | May 28th' A1 | | Gd | 4 M Malonate | statistics on 71 of 106 |
| B21 | 25.01.2009 | 1.48129 | P622 | 49.02 | 49.02 | 303.41 | 0.203 | 2.83 | 3.38 | 27402 | 81.0 [3.02] | 23.0 | 25.0 | 0.25 | 48 [1.25] | | | Gd | | |
| H | 26.02.2009 | 0.97934 | C222 | 55.20 | 55.20 | 153.70 | 0.179 | 3.18 | 4.10 | 15800 | 40.3 [3.54] | 22.6 | 39.6 | 0.72 | 260 [1.0] | Jan27th well B2 | | ? | 30% ethalene glycol | cell dimensios for P3 SG |
| J | 26.02.2009 | 0.97934 | P6222 | 52.81 | 52.81 | 149.66 | 0.401 | 2.55 | 3.00 | 95932 | 64.1 [2.65] | 24.2 | 17.4 | 0.22 | 360 [1.0] | Jan27th well B2 | | ? | Malonate | |
| K1 | 27.02.2009 | 0.97934 | P6222 | 49.49 | 49.49 | 151.66 | 0.377 | 2.68 | 3.15 | 96650 | 87.9 [2.86] | 21.2 | 12.3 | 0.30 | 360 [1.0] | Jan27th A1 well | | SeMet | 70% Malonate (2.8 M) | streaky spots |
| J2 | 28.02.2009 | peak | P622 | | | | | | | | | | | | 120 [1.0] | Jan27th B4 | | SeMet | 70% Malonate (2.8 M) | |
| | inflection | | P622 | | | | | | | | | | | | 120 [1.0] | Jan27th B4 | | SeMet | 70% Malonate (2.8 M) | |
| | remote | 0.97934 | P622 | 49.73 | 49.73 | 151.08 | 0.331 | 2.75 | 3.30 | 91150 | 85.8 [3.09] | 27.2 | 14.1 | 0.22 | 360 [1.0] | Jan27th B4 | | SeMet | 70% Malonate (2.8 M) | |
| J3 | 28.02.2009 | 0.97934 | P6222 | 49.70 | 49.70 | 151.28 | 0.339 | 2.71 | 3.45 | 34855 | 84.8 [3.10] | 22.1 | 23.6 | 0.29 | 360 [1.0] | Jan27th B4 | | SeMet | 70% Malonate (2.8 M) | statistics on 139 of 360 images |
| J7 | 28.02.2009 | 0.97934 | P6222 | 49.83 | 49.83 | 151.43 | 0.396 | 2.94 | 3.30 | 41653 | 84.7 [3.17] | 18.9 | 13.2 | 0.21 | 120 [1.5] | Jan27th B4 | | SeMet | 70% Malonate (2.8 M) | |
| J8 | 28.02.2009 | 0.97934 | P6222 | 52.66 | 52.66 | 150.95 | 0.403 | <3.18 | <3.18 | 25671 | 70.5 [3.18] | 13.5 | 11.7 | 0.18 | 90 [2.0] | Jan27th B4 | | SeMet | 70% Malonate (2.8 M) | |
| A442 | 22.03.2009 | 0.97949 | P6222 | 49.49 | 49.49 | 150.76 | 0.496 | 2.63 | 3.05 | 106272 | 90.9 [2.94] | 13.6 | 7.2 | 0.22 | 180 [2.0] | | | SeMet | | |
| | inflection | | P6222 | | | | | | | | | | | | | | | SeMet | | |
| A4A | 22.03.2009 | 0.97952 | P6222 | 49.76 | 49.76 | 151.81 | 0.589 | 2.37 | 2.79 | 67083 | 86.1 [2.53] | 9.7 | 10.3 | 0.19 | 90 [2.0] | | | SeMet | | |
| | inflection | 0.97965 | P6222 | 49.82 | 49.82 | 151.87 | 0.574 | 2.40 | 2.92 | 65574 | 88.5 [2.73] | 9.9 | 9.1 | 0.19 | 90 [2.0] | | | SeMet | | |
| | remote | 0.97024 | P6222 | | | | | | | | | | | | 90 [2.0] | | | SeMet | | |
| A4B | 22.03.2009 | 0.97934 | P622 | 49.22 | 49.22 | 150.90 | 0.557 | <2.50 | 2.87 | 44287 | 85.7 [2.72] | 11.5 | 11.0 | 0.20 | 60 [2.0] | | | SeMet | | |
| | peak | 0.97949 | P6222 | 49.30 | 49.30 | 150.99 | 0.550 | 2.30 | 2.85 | 65820 | 86.8 [2.72] | 12.1 | 10.5 | 0.20 | 90 [2.0] | | | SeMet | | |
| | inflection | | P6222 | | | | | | | | | | | | 90 [2.0] | | | SeMet | | |
| | remote | 0.97022 | P6222 | 49.42 | 49.42 | 151.20 | 0.499 | 2.50 | 3.05 | 65275 | 83.4 [2.69] | 16.4 | 13.5 | 0.21 | 90 [2.0] | | | SeMet | | |
| A4native | 22.03.2009 | 0.97934 | P622 | 48.05 | 48.05 | 151.10 | 0.353 | <3.25 | 2.90 | 34470 | 72.8 [2.87] | 15.7 | 10.0 | 0.15 | 90 [2.0] | | | Native | | wrong cell low res |
| | high res | 0.97934 | P622 | 48.15 | 48.15 | 302.70 | 0.385 | 1.97 | 2.40 | 407073 | 83.5 [2.23] | 31.4 | 16.5 | 0.22 | 180 [2.0] | | | Native | 30%:3350:16%:400 | high res correct cell |
| tet 1c | 28.08.2009 | 0.97948 | P622 | 51.61 | 51.61 | 150.80 | 0.379 | 2.60 | 3.00 | 55958 | 73.6 [2.76] | 27.5 | 25.2 | 0.21 | 120 [1.5] | cold room 1.25 M AmSo ?buffer? | | Native | 18%/Gly / 1M AmSo | |
| tet bd26 | 28.08.2009 | 0.97948 | | 52.35 | 52.35 | 149.31 | ? | ? | ? | ? | ? [?.?] | ? | ? | ? | 180 [2.0] | cold room 1.25 M AmSo ?buffer? | | Native | 18%/Gly / 1M AmSo | |
| tet bd_a4 | 28.08.2009 | 0.97948 | P6222 | 49.32 | 49.32 | 150.59 | 0.539 | 2.80 | 3.25 | 40792 | 84.5 [2.99] | 13.4 | 14.7 | 0.15 | 90 [2.0] | cold room 1.25 M AmSo ?buffer? | | Native | 3.3M AmSo 5%Gly | |
| tet shA | 28.08.2009 | 0.97948 | P6222 | 50.20 | 50.20 | 151.13 | 0.363 | 2.75 | 3.24 | 38082 | 69.8 [3.07] | 21.2 | 13.9 | 0.22 | 90 [2.0] | cold room 1.25 M AmSo ?buffer? | | Native | 70% Malonate (2.8 M) | |
| jan7b6 | 29.08.2009 | 0.97931 | P6222 | 51.97 | 51.97 | 149.86 | 0.475 | 2.55 | 2.95 | 53702 | 77.6 [2.75] | 12.7 | 18.6 | 0.24 | 180 [1.0] | 0.9M AmSo 50 mM MES 6.5 | | SeMet | 3.3M AmSo 5%Gly | |
| | inflection | 0.97947 | P6222 | 52.40 | 52.40 | 150.58 | 0.446 | 2.65 | 3.09 | 52172 | 74.3 [2.76] | 13.8 | 23.0 | 0.27 | 180 [1.0] | | | | | |
| jan7c3a | 29.08.2009 | 0.97939 | P6222 | 53.08 | 53.08 | 149.55 | 0.385 | 2.85 | 3.50 | 38523 | 68.8 [3.26] | 17.2 | 13.0 | 0.27 | 90 [2.0] | 1M AmSo 50 mM MES 6.5 2.5% PEG 400 | | SeMet | ? 3.3M AmSo 5%Gly ? | |
| | inflection | | | | | | 0.426 | | | | | | | | | | | | | |
| long1b | 29.08.2009 | 0.97948 | P6222 | 49.51 | 49.51 | 152.62 | 0.244 | 3.03 | 3.80 | 81759 | 82.1 [3.31] | 23.9 | 16.4 | 0.58 | 240 [1.5] | 1.1 M AmSo NaAcetate?? | | Native | 3.3M AmSo 6%Gly | |
| long1e | 29.08.2009 | 0.97948 | P6222 | 51.77 | 51.77 | 150.11 | 0.513 | 2.35 | 2.92 | 54193 | 67.6 [2.57] | 12.2 | 16.0 | 0.29 | 180 [1.0] | 1.1 M AmSo NaAcetate?? | | Native | 10 µl mother liquer + 1 µl 33%3350/6%400 | |
| phb4 | 29.08.2009 | 0.97934 | P6222 | 49.93 | 49.93 | 151.51 | 0.493 | 2.55 | 3.20 | 56507 | 85.4 [2.88] | 11.9 | 14.9 | 0.31 | 180 [1.0] | pH clear a#' 0.8 M AmSo buffer? pH 7 | | SeMet | 3.3M AmSo 6%Gly ??? | |
| | inflection | 0.97947 | P6222 | 49.93 | 49.93 | 151.33 | 0.312 | 2.87 | 3.68 | 50533 | 80.5 [3.14] | 25.2 | 26.1 | 0.35 | 180 [1.0] | pH clear a#' 0.8 M AmSo buffer? pH 7 | | SeMet | 3.3M AmSo 6%Gly ??? | |
| IX a | 01.06.2010 | 0.9795 | P6122 | 48.00 | 48.00 | 301.30 | 0.69 | <1.98 | 2.00 | 192458 | 40.2 [1.98] | 9.6 | 7.3 | 0.23 | 360 [0.5] | | | Native | | |
| IX lowres | 01.06.2010 | 0.9795 | P1 | 48.10 | 48.00 | 300.70 | 0.78 | 93.43 | Å | 39008 | 10.6 [3.43] | 2.9 | 4.0 | 0.26 | 360 [0.5] | | | Native | | |
| IIX1 | 01.06.2010 | 0.9795 | P6122 | 47.50 | 47.50 | 298.40 | 0.69 | 2 | 2.33 | 175028 | 29.3 [1.98] | 9.0 | 14.0 | 0.20 | 360 [0.5] | | | Native | | |
| one_a | 01.06.2010 | 0.9795 | C2 | 49.60 | 49.40 | 309.70 | 0.66 | 2.5 | 2.95 | 105520 | 41.9 [2.45] | 11.7 | 15.4 | 0.34 | 374 [0.5] | | | Native | | |
| two | 01.06.2010 | 0.9795 | P6122 | 48.60 | 48.60 | 302.50 | 0.68 | <2.34 | 2.60 | 118682 | 62.3 [2.34] | 8.4 | 8.8 | 0.23 | 587 [0.25] | | | Native | | |
| three | 01.06.2010 | 0.9795 | P6122 | 48.90 | 48.90 | 306.10 | 0.67 | 2.48 | 2.68 | 112660 | 72.6 [2.48] | 13.4 | 9.5 | 0.29 | 597 [0.3] | | | Native | | |
| four | 01.06.2010 | 0.9795 | P1 | 48.50 | 48.50 | 302.10 | 0.69 | 1.98 | 2.60 | 179313 | 8.5 [1.98] | 6.0 | 21.2 | 0.26 | 360 [0.5] | | | Native | | |
| fourlowres | 01.06.2010 | 0.9795 | C222 | 49.10 | 48.90 | 303.70 | 0.69 | <2.94 | 3.09 | 35559 | 16.6 [2.94] | 5.4 | 8.9 | 0.30 | 490 [0.2] | | | Native | | (inverse beam) |
| bijv | 01.06.2010 | | P6222 | 49.00 | 49.00 | 151.10 | 0.65 | <3.27 | 3.50 | 30625 | 83 [3.27] | 16.5 | 14.2 | 0.45 | 346 [0.5] | | | SeMet | | |

## 8.3 Appendix-3 Diffraction Images

SeMet-Small

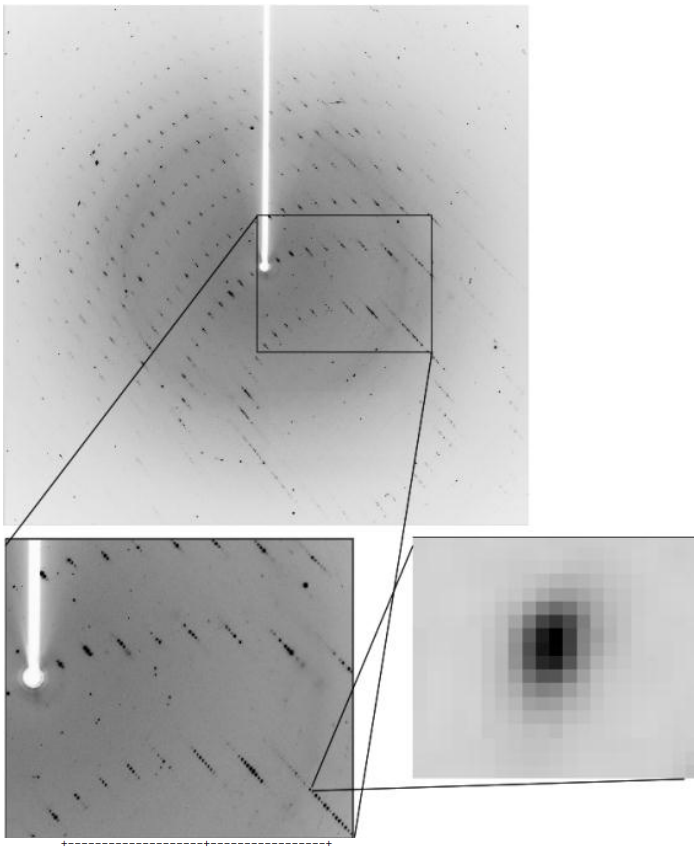


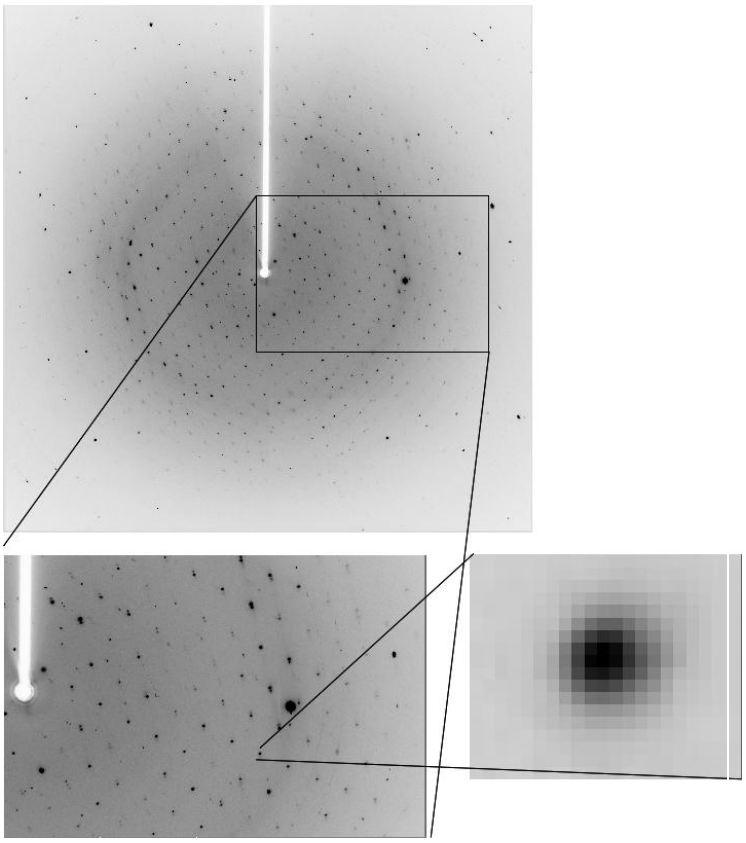


```
Selected Group is:    P6(2)22,  #180

Unit Cell:      49.51   49.51  151.10   90.00   90.00  120.00

--- INTEGRATION ---

--- Summary ---
Observed Reflections:       67109
Unique Reflections:          4505
Redundancy:    14.9
Unit Cell:    49.76   49.76  151.81   90.00   90.00  120.00
Cell E.S.D:     0.1     0.1    0.29       0       0       0
Mosaicity:     0.19
Standard deviation of spot position:    1.220 (pix)
Standard deviation of spindle position: 0.230 (deg)

--- Statistics ---

   Shell Complete   R_meas R_mrgd-F  I/Sigma   SigAno  AnoCorr
    6.38   94.00%    5.70%    1.80%    44.49     3.54     88.0
    4.64   99.80%    6.30%    2.20%    39.84     2.77     85.0
    3.82   97.20%    7.20%    2.90%    35.42     2.08     75.0
    3.32   95.10%   15.70%    8.00%    18.07     1.34     47.0
    2.98   89.50%   42.80%   21.80%     7.96     0.93     18.0
    2.73   78.50%  108.50%   57.10%     3.25     0.82      4.0
    2.53   64.10%  174.00%  115.30%     1.57     0.76     -6.0
    2.37   57.80%  213.80%  176.20%     1.00     0.74      6.0
    2.23   45.00%  176.50%  145.90%     0.98     0.67    -24.0
   Total   75.60%   10.20%   16.60%    15.35     1.32     42.0

Resolution cut-off from preliminary analysis (I/SigI>1.5):
2.53
```

```
- SCALING ---

- Statistics ---

 Shell Complete   R_meas R_mrgd-F  I/Sigma   SigAno  AnoCorr
 11.32   76.20%    5.20%    1.90%    46.97     4.32     93.0
  8.00   95.90%    3.40%    1.10%    47.82     4.34     97.0
  6.53   94.70%    4.90%    1.80%    43.38     3.70     83.0
  5.66   97.30%    5.00%    1.90%    40.26     3.78     87.0
  5.06   99.60%    5.30%    2.10%    36.74     3.17     84.0
  4.62   99.30%    4.80%    2.30%    37.70     2.84     83.0
  4.28   99.70%    5.10%    2.30%    38.27     2.38     77.0
  4.00   96.60%    6.20%    3.00%    31.84     2.46     81.0
  3.77   97.10%    9.30%    4.90%    23.19     1.87     68.0
  3.58   93.50%   12.80%    6.90%    18.24     1.54     54.0
  3.41   93.10%   17.60%   10.40%    13.30     1.31     42.0
  3.27   94.60%   22.00%   16.10%    11.09     1.11     32.0
  3.14   91.70%   36.30%   18.90%     7.80     0.96     11.0
  3.02   84.40%   55.50%   38.80%     4.91     0.86      5.0
  2.92   81.20%   75.50%   50.40%     3.89     0.90      9.0
  2.83   77.80%   99.40%   69.70%     2.84     0.81      1.0
  2.74   73.20%  124.90%   76.80%     2.21     0.79     -4.0
  2.67   68.50%  143.00%  106.80%     1.66     0.71     -5.0
  2.60   63.20%  155.00%  146.00%     1.38     0.79     -9.0
  2.53   56.30%  182.10%  170.80%     1.13     0.70    -17.0
 Total   84.00%    8.50%   13.20%    16.77     1.53     47.0
```

Nat-Large





| Dataset | a |
|---|---|
| Score [a] | 0.69 |
| Wavelength (A) | 0.9795 |
| Space Group [b] | P6(1)22 |
| Unit Cell (A) | 48.0 48.0 301.3 |
| (deg) | 90.0 90.0 120.0 |
| Cell Volume (A^3) | 599939 |
| Resolution (A)[c] | 1.98 |
| All Reflections | 192458 |
| Unique Reflections | 13899 |
| Multiplicity | 13.8 |
| Completeness | 89.8% |
| Mosaicity | 0.23 |
| I/sigma(I) [d] | 20.9 |
| R-mrgd-F [e] | 7.3% |
| R-meas [f] | 9.6% |
| sigma(spot) (pix) | 1.5 |
| sigma(angle) (deg) | 0.1 |
| No. Ice rings | 5 |

Correction Statistics

| Resolution | Completeness | R_meas | R_mrgd-F | I/Sigma | SigAno | AnoCorr |
|---|---|---|---|---|---|---|
| 5.88 | 95.2 | 5.3 | 1.8 | 46.38 | 0.544 | -34 |
| 4.18 | 99.9 | 5.8 | 1.8 | 47.93 | 0.536 | -42 |
| 3.42 | 100.0 | 7.4 | 2.1 | 43.5 | 0.642 | -28 |
| 2.96 | 99.9 | 12.3 | 3.0 | 32.16 | 0.679 | -23 |
| 2.65 | 99.7 | 23.3 | 4.6 | 21.59 | 0.636 | -11 |
| 2.42 | 98.9 | 38.4 | 8.6 | 12.96 | 0.622 | -11 |
| 2.24 | 95.0 | 44.7 | 18.9 | 8.02 | 0.693 | -7 |
| 2.10 | 84.6 | 67.0 | 43.7 | 4.35 | 0.678 | -9 |
| 1.98 | 53.0 | 39.8 | 47.6 | 3.3 | 0.689 | -15 |
| total | 89.3 | 9.7 | 7.2 | 21.16 | 0.646 | -15 |

Preliminary resolution cut-off (Based on I/sigma(I) > 1 cut-off.): 1.98

Scaling Statistics

Statistics presented by resolution shell

| Resolution | Completeness | R meas | R mrgd-F | I/Sigma | SigAno | AnoCorr |
|---|---|---|---|---|---|---|
| 8.86 | 86.0 | 4.9 | 1.8 | 42.6 | 0.468 | -29 |
| 6.26 | 99.7 | 5.3 | 1.7 | 45.17 | 0.525 | -44 |
| 5.11 | 100.0 | 6.0 | 1.9 | 44.47 | 0.546 | -39 |
| 4.43 | 99.8 | 5.6 | 1.7 | 46.34 | 0.494 | -43 |
| 3.96 | 100.0 | 6.1 | 1.8 | 45.77 | 0.573 | -36 |
| 3.62 | 100.0 | 7.2 | 2.1 | 42.97 | 0.64 | -28 |
| 3.35 | 99.9 | 8.7 | 2.1 | 39.24 | 0.638 | -30 |
| 3.13 | 100.0 | 10.9 | 2.6 | 33.58 | 0.678 | -20 |
| 2.95 | 99.9 | 15.5 | 3.5 | 27.76 | 0.691 | -21 |
| 2.80 | 99.9 | 19.0 | 3.7 | 24.38 | 0.623 | -13 |
| 2.67 | 99.6 | 27.7 | 4.9 | 19.94 | 0.668 | -12 |
| 2.56 | 99.2 | 36.4 | 7.4 | 15.38 | 0.691 | -11 |
| 2.46 | 99.0 | 39.1 | 8.7 | 12.58 | 0.64 | -8 |
| 2.37 | 98.0 | 41.1 | 12.8 | 10.31 | 0.62 | -15 |
| 2.29 | 97.2 | 43.3 | 18.9 | 7.72 | 0.669 | -10 |
| 2.21 | 85.5 | 56.9 | 28.6 | 7.01 | 0.787 | -10 |
| 2.15 | 88.1 | 76.6 | 49.1 | 4.57 | 0.75 | -4 |
| 2.09 | 81.4 | 43.3 | 46.2 | 3.91 | 0.644 | -10 |
| 2.03 | 65.9 | 42.6 | 47.8 | 3.66 | 0.711 | -16 |
| 1.98 | 40.2 | 31.2 | 45.8 | 2.86 | 0.666 | -12 |
| total | 89.8 | 9.6 | 7.3 | 20.85 | 0.655 | -15 |

## 8.4 Appendix-4 Possible Twin Operators

Possible twin operators for trigonal and hexagonal space groups

| Space-group number | Space group | Point group | Possible twin operator |
|---|---|---|---|
| 143 | *P*3 | PG3 | *-h, -k, l* or *k, h, -l* or *-k, -h, -l* |
| 144 | $P3_1$ | PG3 | *-h, -k, l* or *k, h, -l* or *-k, -h, -l* |
| 145 | $P3_2$ | PG3 | *-h, -k, l* or *k, h, -l* or *-k, -h, -l* |
| 146 | *R*3 | PG3 | *k, h, -l* |
| 149 | *P*312 | PG312 | *-h, -k, l* or *k, h, -l* |
| 151 | $P3_112$ | PG312 | *-h, -k, l* or *k, h, -l* |
| 153 | $P3_212$ | PG312 | *-h, -k, l* or *k, h, -l* |
| 150 | *P*321 | PG321 | *-h, -k, l* or *-k, -h, -l* |
| 152 | $P3_121$ | PG321 | *-h, -k, l* or *-k, -h, -l* |
| 154 | $P3_221$ | PG321 | *-h, -k, l* or *-k, -h, -l* |
| 168 | *P*6 | PG6 | *k, h, -l* |
| 169 | $P6_1$ | PG6 | *k, h, -l* |
| 170 | $P6_5$ | PG6 | *k, h, -l* |
| 171 | $P6_2$ | PG6 | *k, h, -l* |
| 172 | $P6_4$ | PG6 | *k, h, -l* |
| 173 | $P6_3$ | PG6 | *k, h, -l* |
| 177 | *P*622 | PG622 | *k, h, -l* |
| 178 | $P6_122$ | PG622 | None |
| 179 | $P6_522$ | PG622 | None |
| 180 | $P6_222$ | PG622 | None |
| 181 | $P6_422$ | PG622 | None |
| 182 | $P6_322$ | PG622 | None |